

Taxonomy of Directing Semantics for Film Shot Classification

Hee Lin Wang and Loong-Fah Cheong

Abstract—The immense indexing potential of motion cues has hitherto been realized only in domains with more apparent structure (e.g., sport videos). To address the lack of theoretical attention and to realize the potential of motion-based indexing in the subtler film domain, we propose a systematic approach to build taxonomy for film directing semantics. These motion-related semantics are grounded upon cinematography and are thus more appealing to users. In order to automate the classification of these semantics, we have developed a novel markov random field based motion segmentation algorithm with an integral foreground/background identification capability based on edge occlusion reasoning. This algorithm is sufficiently robust and fast for film domain conditions, and allows us to formulate salient and novel motion descriptors capable of mapping to the proposed directing semantics. We demonstrate the validity of the framework and effectiveness of the motion-based descriptors by classifying shots from Hollywood domain movies according to the proposed taxonomy with satisfactory results.

Index Terms—Attention, classification, film directing semantics, film grammar motion, movie shot.

I. INTRODUCTION

IN CONTRAST to other types of media, the defining aspect of film is the narrating of stories through the use of camera motion. It is often through motion that the content or the meaning in a shot is expressed and the attention of the viewers captivated or shifted, allowing the film's intentions to be communicated. Due to the widespread use of motion in connoting certain cinematic meaning and narrative flows in video genres like film, sports, and program shows, motion has gradually gained recognition as a potent feature for video indexing. This is well attested by the inclusion of motion descriptors in the MPEG-7 [1] standard such as motion activity, camera movement, trajectory, and parametric motion, underlining the immense potential of motion cues for semantic indexing of film shots.

Most prior works on motion-based indexing are limited to the sports domain [2]–[4], whose organized structure allows easier application of motion cues to recover semantics. In contrast, the structure of the narrative video genre (e.g., film)

is much more varied and tends to be embedded with subtler forms of semantics. Thus attempts to extract film semantics using simple low level motion-based features are met with limited success. In particular, these efforts are hampered by two shortcomings: lack of a framework that can effectively integrate the wealth of cinematic-domain knowledge; and furnish more meaningful film shot categories; and a lack of well-founded intermediate-level motion features that capture camera works routinely accomplished by the cameraman or the producer and thus facilitate meaningful classification.

This paper addresses the first issue by exploiting the underlying relationship that exists between film shot semantics and motion features. In cinematography, a widely accepted set of informal production rules often governs the relationship between the meanings of the film shots to be conveyed and various camera-related attributes, especially camera motion. For instance, tracking operations in a shot make intentionality manifest; they very likely indicate the presence of subject(s) of interest, which is of strong indexing value. Another common directing rule is that of varying the camera distance from the subject(s) of interest to subtly adjust the relative emphasis between the subject and the surrounding environment in accordance to the director's intentions. Exploiting these rules, which are also known as film grammar, enables us to propose a cinematographically grounded taxonomy of shot semantics that are of sufficiently high level to be of interest to the users. The approach we use to build this taxonomy is scalable to algorithmic advances and suggests the most comprehensive directing semantics taxonomy possible for a given level of ability to compute directing descriptors.

On the second issue, we develop a novel markov random field (MRF) based algorithm for motion segmentation. Foreground/background labeling is an integral part of this segmentation process; the labeling is achieved with the help of occlusion reasoning, as opposed to often *ad hoc* steps in other MRF works. The segmentation process is also guided by cinematic-domain knowledge, based upon how viewer attention is usually manipulated in the cinema domain. Given this foreground/background labeling, we subsequently introduce an attention descriptor that effectively tracks the evolution of viewer's attention. The resultant system is demonstrated on a large video corpus to obtain satisfactory experimental results.

We focus on Hollywood films, as they contain a full range of cinematographic motions prescribed by directing grammar. Nevertheless, directing grammar is also employed in the production of the vast majority of fictional video narratives ranging from dramas to mini series; thus the application arena

Manuscript received March 3, 2008; revised July 7, 2008 and October 10, 2008. First version published May 12, 2009; current version published September 30, 2009. This paper was recommended by Associate Editor J. Luo.

H. L. Wang is with the Institute of Infocomm Research, Agency for Science, Technology and Research, Singapore 138632 (e-mail: hlwang@i2r.a-star.edu.sg).

L.-F. Cheong is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576 (e-mail: eleclf@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCSVT.2009.2022705

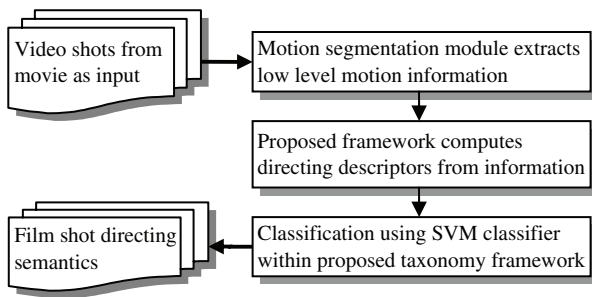


Fig. 1. Flowchart of system overview.

of our work is by no means restricted to Hollywood films. Since the shot—defined as a continuous and uninterrupted series of frames from one camera—has been suggested as the basic unit of meaning in film content analysis, we seek to recover some aspect of this atomic meaning at the shot level.

To our knowledge, no work on the film domain proposes and recovers such a comprehensive set of directing shot semantics, explicitly drawn from film grammar, as ours. Though motion is the modality under investigation, our framework readily allows other modalities or even forms of semantics (e.g., affective) to complement and enhance the indexing capabilities of the proposed film shot semantics.

We envisage that our semantic-level shot classification can be used to drive applications such as automated film analysis [5], editing [6], film structure creation [7], retrieval [8], and video preview/summarization [9]. For instance, it allows interesting foreground objects in tracking shots to be detected and retrieved (e.g., favorite actor/actress). Users who wish to create location-based “albums” of scenic movies may wish to select only panning background shots. Furthermore, leveraging on the directing “blueprint” underlying the shots, various flavors of movie previews (e.g., a smoother or a more dynamic style) can be created depending on camera/object motion. Finally, movie structure analysis, in conjunction with a set of tailored rules to detect specific semantic patterns, may allow us to extract the more significant/interesting parts of a movie and splice them into coherent summaries.

The rest of the paper is as follows. In Section II, we exploit cinematographic domain knowledge to develop a motion-based film shot indexing framework. Section III proposes a novel motion segmentation technique, specially adapted to facilitate the extraction of salient features for film shot semantics recovery. Section IV elaborates further upon the motion-based shot features derived under the framework, as well as the probabilistic inference engine used to infer the shot directing semantics. Experimental results are presented in Section V, followed by the conclusion in Section VI. The system overview is illustrated in Fig. 1.

II. MOTION-BASED INDEXING FRAMEWORK

In comparison to the well-structured directing format for sports domain videos, shot semantics arising from the film domain, at least at a sufficiently high and interesting level, are far more complex. In this section, we will propose taxonomy of film directing semantics that is motivated by the current cinematographic conventions and organized around several key

film directing devices. We also show that by carefully exploiting the constraint afforded by the cinematic environment, these semantic classes can be obtained directly from various motion-based features without going through the difficult process of full camera motion recovery.

A. Literature Review

Motion-based query systems can be classified according to whether they are primarily designed for retrieval or for indexing. The former seeks to locate items that are similar to the provided examples, and their frameworks are generally not optimized with any domain in mind. In contrast, motion-based indexing exploits *a priori* information from specific domain so as to formulate semantic classes and models for classification.

For motion-based retrieval works, Idris [10], Oh [11], and Fablet [12] used frame and shot level motion features for shot classification, without performing any motion segmentation. Some teams in TRECVID 2005 [13] carried out simple shot-level camera motion classification into pan, tilt, and zoom, based upon motion vectors provided in the MPEG stream. It should be noted that these vectors are meant to minimize inter-frame differences and do not necessarily conform to the true optical flow which our algorithm attempts to recover. More recently, Ho [14] computed similarity measures based on the motion within “regions of interest.” However, it is difficult to extend such work in the total absence of the object concept. Among the works that feature explicit motion segmentation, Mezaris [15], Dagtas *et al.* [8] with PICTURESQUE and Chang *et al.* [16] with VideoQ used various impressive motion-based descriptors for retrieval. However, these descriptors do not exploit film grammar to recover directing semantics.

Contrary to the retrieval works, indexing works explicitly map motion features to high-level semantics. Takagi [4] used the statistical properties of the transition between camera motion types (pan, tilt, zoom, shake) to distinguish between the different genres of sports footages (baseball, soccer, tennis, sumo wrestling, etc.). Lazarescu [3] derived discrete intermediate descriptors such as camera angle, speed, number of stages in camera motion, and net pan/tilt to recognize different parts of offensive plays in American football. Ma [17] employed a circular polar representation of the optical flow to compute statistical measures for classifying object shots (with objects), camera shots (no objects), and finally non-semantic shots (no meaningful movement). Highly structured domains such as sports are usually the subject of such works to ensure the feasibility of defining a “vocabulary” of actions whose computable features can map to semantics, in contrast to the paucity of works [18] involved in the subtler film domain.

In contrast, a taxonomy explicitly based on the production rules underlying a movie enjoys greater relevance in film domain applications. Take for instance movie summarization. Without the knowledge of cinematography, a movie may be blindly summarized by maximizing its visual entropy. However, with cinematic-domain-based taxonomy, it is possible to detect the more significant/interesting parts of a movie from how directors direct viewer attention and splice these parts into a summary. For instance, if two shots with decreasing camera

TABLE I

DIRECTING SEMANTICS ORGANIZATION BY FILM DIRECTING ELEMENTS

Semantics	Camera motion/ FOA behavior	Camera distance
1) Stationary: most dialogue and close-ups fall into this category	Little camera and FOA motion	NA: camera distance not recovered due to weak motion
2) Contextual-Tracking: indexes FOA while showing its relation to the surroundings	Usually smooth camera following a particular FOA	Long
3) Focus-Tracking: indexes interesting FOA by trailing object closely		Close-Up/ Medium
4) Focus-In: used to increase tension and sense of emotional involvement	Camera behavior marked mainly by decreasing camera distance	Decreasing
5) Focus-Out: to pull back and relocate or detach viewer emotionally	Camera behavior marked mainly by increasing camera distance	Increasing
6) Intermittent/Panning Establishment: Indexes interesting locales or surroundings. Used to introduce locations in plot	Smooth camera motion with no FOA	Not Applicable: absence of FOA
7) Chaotic: usually indexes fast action shots where high motion encounters occur	Large and un-patterned movements of both FOA and camera	Not Applicable: Incoherent FOA and camera motion

distance are shown together, it is a clear-cut sign of focusing on interesting objects. Similarly, tracking shots are another trustworthy sign of the presence of interesting objects.

B. Developing the Framework

Film directing grammar—henceforth termed loosely and interchangeably as “directing”—is one of the most crucial set of production rules underlying the movie making process. It facilitates conveying the director’s intentions through specific camera motions and viewpoint attributes. By so doing, viewer attention [9] is directed to concentrate on an object or place [the focus of attention (FOA)], thus privileging one set of viewer observation, interpretation, and experience. It is this act of attention directing that helps define the semantics of a given shot in three major ways.

- 1) Camera motion operations: Various camera motions are routinely executed to convey director’s intentions. Some of these routines include the introduction to a scene, the unfolding of an event, the tracking of an object and the directing of attention. As we have mentioned, tracking operations strongly imply the presence of interesting FOAs.
- 2) Camera distance from FOA: The framing distance alone can evoke substantially different subjective response, an example being that close-ups tend to have more emotional impact than long distance shots.
- 3) On-screen duration of FOA: The perception of the passage of time can be easily manipulated via changes

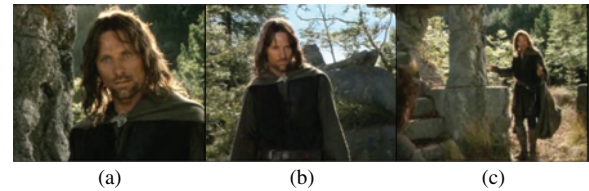


Fig. 2. Example shots at different camera distances. Typical images of (a) close-up, (b) medium, and (c) long shot.

in the pace with which camera viewpoint and motion are switched or executed.

Accordingly, the contribution of this paper is to analyze how film directing grammar allows us to compute suitable directing elements, in order to extract their corresponding shot semantics. In the following, we introduce the directing elements in greater detail and show how to construct taxonomy of directing semantics computable from these directing elements.

1) *Camera Motion/FOA Behavior*: Traditionally, the major camera motion types are pan (rotation about the vertical axis), tilt (rotation about the horizontal axis), and finally zoom (change in focal length). For the purpose of semantic level video indexing, it is not so much the exact amount of zoom but the presence of such camera operation and the qualitative degree of movement that are important. Thus we are not concerned with the quantitative aspect of the camera motion.

The type of camera motion in tandem with the FOA behavior conveys far more meaning than most consciously realize, just like we unconsciously inhabit the personal space of characters in a film. In the simplest case, the stationary shot can be used to signify a calm scene or a pause pregnant with meaning. The tracking and establishment shots often have similar camera motion type, but their semantics are very different. These two types of shots can be separated based on whether the shots are following an FOA or not. Yet another major motion type with strong semantic significance is the zoom-in and, to a lesser extent, the zoom-out motion; the impact of such shot lies not so much in its cultural perception but stems directly from its psychological relevance: human beings are responsive to impending colliding objects. Thus, each of these camera motions/FOA behaviors can be related to some high-level semantic information, as summarized in Table I. Finally, it is recognized that not all shots are characterized by coherent camera motion and FOA behavior. Hence we also create the chaotic shot category to describe shots that do not exhibit coherent motion patterns.

2) *Camera Distance*: Camera framing refers to the manner the FOA(s) are presented in the frame, which includes the aspects of distance, composition, angle (low/high angled shot), level (degree of canting), and height [18]. Of these five, we concentrate on one of the most influential aspects of framing: camera distance. Cinematography admits of three coarse distance gradations (Fig. 2): close-up, medium, and long. Because emotional involvement and degree of attention are approximately inversely correlated with camera distance, it can be directly used as a semantic index for the amount of attention and emotional proximity.

For the purpose of indexing, close-ups and medium shots are consolidated into one group for the following two reasons. Firstly, the distinction between close/medium shots is less clear compared to the separation between medium/long shots: there is a continuous gradation from the close-ups to the medium shots due to the varieties of ways of framing a person. Secondly, the camera distance is directly related to the size of the field of view, and serves as a good index of whether the director intends the shot to offer a broad overview or specific focus [20]. This distinction in purpose usually coincides with the demarcation between close/medium and long shots.

C. Taxonomy for Film Directing Semantics

Directing semantics are created by manipulating various directing elements, which can thus be utilized as a basis to organize the semantic taxonomy. To visualize the relationship between the directing elements and the semantics, we generate a table from all possible permutations of the directing elements (i.e., camera motion/FOA behavior and camera distance). Then the most meaningful and frequently employed directing semantics are selected from film grammar and assigned to their corresponding permutations within the table (Table I).

The exhaustive permutations of directing elements ensure that every shot that is pertinent in the directing semantics context can be suitably assigned to one of the classes in the taxonomy. Impossible combinations are easily detected and removed. For instance, the definition of zoom camera behavior precludes any combination with fixed camera distance. At the same time, new semantic classes that contribute to a richer taxonomy appropriate for video indexing may emerge, as in the case of Tracking (Table I, second row), which has been split into the meaningful focus-tracking and contextual-tracking classes according to the camera distance element. Finally, under some circumstances, different semantics with the same permutation of directing elements can be merged together. For instance, shots that do not belong to any of the first six semantic classes are generally characterized by patternless camera/FOA motions. In such a case, a new semantic class with the label “chaotic,” so called due to the incoherent nature of the shots motion-wise, is used to house these shots.

We finally settle on seven semantic classes: 1) static, 2) contextual-tracking (C-Track), 3) focus-tracking (F-Track) 4) focus-in (F-In), 5) focus-out (F-Out), 6) establishment, and 7) chaotic. Although directing semantics in films may be “fuzzier” in reality and less mutually exclusive compared to the clear-cut categories typically found in the sports domain (goal shots, replay, etc.), useful indexing can still be garnered from these semantics, as a fuller description of the semantic classes and their significance in the following paragraphs illustrate.

Establishment shot: In film grammar, scenes or story units usually start with an establishment shot (Fig. 3, second row) [21], [22]. These shots are realized using smooth panning motion to survey the new location, introducing or reminding the viewer of a new environment or spatial relationships inside it. Establishment shot detection aids in story structure recovery such as analyzing the story units of a movie and



Fig. 3. Examples of semantic classes. Contextual-tracking (1st row), establishment (2nd row), focus-in (3rd row), and chaotic (4th row).

scene segmentation. Another closely related common technique in cinematography is the intermittent pan [22], where the camera focuses the attention on multiple—usually two—FOAs by panning from one FOA to another with smooth camera movements. The chief purpose behind the intermittent pan is however related to that of establishment shot—it uses the panning motion to highlight the spatial relationship between various FOAs within the shot, without focusing exclusively on any of them.

Stationary (Static) shot: As the workhorse shot for many occasions, the vast majority of dialogue shots (over-the-shoulder, two-person shots, etc.) and practically all close-up shots fall under this category [20]. The minimal motion content of such shots serves as a good index of the lull portions of a movie in contrast to its climaxes.

Tracking shot: The tracking shot is defined as one where the primary intention is to draw attention to an FOA by having the camera closely following or rotating around the FOA. This draws the viewer into a closer and more intense relationship with the subject [18] by creating the illusion that the viewer is directly present within the scene itself. Consequently, tracking shots are a valuable index for the strong presence of FOAs and first-person points of view. There exist two major variants of tracking shots. The focus-tracking shot concentrates the viewer’s attention on a subject by employing either close or medium camera shot distance. In contrast, the contextual-tracking shot (Fig. 3, first row) uses the long shot to show off the surroundings while accomplishing the dual purpose of tracking the subject. These two types of shots usually intertwine in longer tracking sequences as the directing requirements alternate.

Focus-In (F-In) shot: For our work, focus-in shots are shots where 1) the subject or camera is moving toward the other and 2) there is apparent collision due to zoom-in (Fig. 3, third row). Both these shots are defined together because they have the similar ability to create and amplify emotional empathy

via an enhanced perception of the character expressions or the visceral tension of impending impact (e.g., colliding car). Furthermore, focus-in shots are also employed to highlight important details.

Focus-Out shot: Focus-out shots, also known as detachment shots, emotionally detach or relax the interest of the viewer from the subject. This effect is usually achieved through zooming out or dolly out shots, as the camera gradually moves away from the subject and creates emotional distance. Since this shot widens the field of view, it is also employed to reveal more information about the surroundings [20].

Chaotic shot: The chaotic shot (Fig. 3, fourth row) refers to shots characterized by large degree of FOA movement, usually in conjunction with un-patterned camera motion. This unique motion behavior covers almost all other combinations of motion behavior not covered by other semantic classes. In this shot type, it is not unusual for the fast moving FOA to dominate viewer attention. Such shots usually cluster around the movie climaxes and tend to be more prevalent in movies of the action genre.

D. Shot Labeling

Before going into the technical details of feature extraction, we should say something about how we manually assign ground truth labels to the shots. Often, it is difficult to annotate shot motion because a significant minority of shots has two equally significant directing semantics present within the shot [13]. For instance, a shot may comprise equally long stationary and tracking parts. For such cases, shots deemed to contain two significant directing semantics are assigned dual labels.

Although there is some inevitable subjectivity in the labeling process, the qualitative nature of our class descriptors reduces the need to make fine quantitative judgments. For instance, we do not need to distinguish between pan and tilt, nor gauge if the amount of tilt is significantly more than that of pan etc. To improve objectivity, we have also formulated a set of systematic guidelines that govern the manual assignment of ground truth for the semantic labels of some of the more ambiguous movie shots.

- 1) If there is only a fragment of background framing the “tracked” foreground (the greenery outside in Fig. 4a), then the background is totally discounted and the shot will be not labeled as a tracking shot but perhaps as a static shot if the foreground is relatively stationary.
- 2) Establishment shots that employ object tracking as a means to introduce new scenery to the viewer are labeled as tracking shots. Establishment shots filmed using a totally static shot, which is a technique used infrequently from time to time, are labeled as static shots (Fig. 4b). This is in recognition of the fact that some forms of establishment shot require more than just motion features to detect and thus lie outside the scope of our work.
- 3) If a stationary camera portrays a long shot, then regardless of the FOA motion, the shot is labeled as a static shot. This is because the foreground area is simply too small to override the effects of the stationary camera (e.g., tiny fragment of FOA in the middle of Fig. 4c).

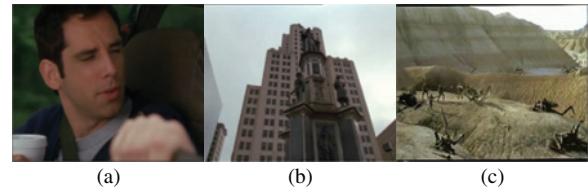


Fig. 4. Example shots to illustrate labeling rules.

III. MOTION SEGMENTATION

In this section, we propose an extension of the MRF-based motion segmentation algorithm by Tsaig [35] to extract motion-related shot descriptors for higher level shot indexing. The extension incorporates both occlusion-reasoning and cinematographic constraints, and is modified for fast, robust, and more accurate fore-ground/background segmentation used by the shot descriptors in the film environment.

A. Literature Review

Motion segmentation can be delineated into two major approaches. The first approach, also referred to as the top-down approach, seeks to iteratively recover the dominant motion from the residue pixels that do not conform to the dominant motion of all previous iterations. Odobez [23] exemplifies this approach, which returns excellent performance especially for the first dominant motion.

The second approach estimates all motions and their respective supports simultaneously, and in turn has three major variants. The first variant, also known as the bottom-up approach, estimates a large number of motions, one for each small image area, before merging patches that are similar in motion [24]–[27]. This approach is attractive both in terms of computation and flexibility, since redundant computation is avoided and the number of motions can vary as circumstances demand. However, the criterion of merging remains problematic. Shi and Malik [28] used a global graph partitioning method known as the normalized cut to best partition pixels into dissimilar sets instead.

The second variant is the level set formulation where the boundaries of curves evolve according to certain partial differential equations [29], [30]. However, the final solution does depend on initialization conditions. The last variant is the expectation maximization (EM) framework [31]–[34], whose critical strength is to allow the best motion hypotheses to explain the observation through competition for support regions. Tweed and Calway [27], Drummond [31], and Torr *et al.* [33] have gone further by integrating some form of occlusion detection mechanism into the segmentation process itself. However, like almost all EM methods, they suffer from the deeply unrealistic requirement to have a fixed and known number of hypothesized motions *a priori*.

B. Our Motion Approach

To decipher shot semantics, it is critical to track the evolution of viewer attention by identifying and tracking both the FOA and the background areas through the entire shot via motion segmentation. Tsaig [34] introduced a simplified fast

and robust MRF-based foreground/background segmentation algorithm that can handle the dynamic film environment by assuming one dominant motion with unknown number of other motions, which is a typical scenario for movie shots. The MRF is a graphical model that falls under the EM framework and offers a formal approach capable of easily modeling a variety of dynamics. However Tsai makes the somewhat broad assumption that the dominant motion always belongs to the background. Such assumption is violated by a small but still significant number of close-up shots.

We have extended Tsai's work chiefly by integrating occlusion reasoning into the motion segmentation process. This allows us to identify the true FOA through inferring the relative depth order between the dominant and nondominant motion areas, without making the aforementioned assumption by Tsai. To our knowledge, this is the first MRF-based paper with integrated occlusion handling mechanism that does not need to know nor fix the number of hypothesized motions.

1) *Spatial Region Segmentation and Motion Estimation*: To improve computation time and segmentation accuracy, motion segmentation is carried out only between consecutive frames which exhibit movement above a preset threshold (intensity differences larger than 6 for more than 5% of the pixels). This is especially useful in dealing with shots with long periods of stationarity. These frames are transformed into a region-based representation which groups spatially connected pixels by intensity into regions using the watershed segmentation algorithm [35].

2) *Hierarchical Optical Flow Estimation*: Next, we utilize a multiscale affine motion model estimation scheme [23], which is known for its good tradeoff between complexity and accuracy, to recover optical flow. It computes flow on the luminance component of the $YCbCr$ color space, and uses the Tukey biweight estimator to limit the deleterious influence of outliers caused by noise, occlusion, and violation of the motion model. This scheme recovers the dominant motion and assigns to every pixel i a weight w_i , whose range is between $[0, 1]$, which can be interpreted as the likelihood of the pixel being assigned the computed optical flow.

Initially, motion estimation is performed over all edge pixels to obtain the dominant motion. This translates to computation time savings of around one-third compared to using all the pixels. Let the average weight of the edge pixels of region i under the current motion estimate be $w(i)$; then all regions with $w(i) > 0.7$ are assigned to the current motion model and their pixels are retired from subsequent motion estimation process. This first recovered motion is designated as the dominant motion and the process is iterated repeatedly for other secondary major motions until the region areas under these secondary motions fall below 15% of the image area.

Amongst the remaining regions not yet assigned, a motion model due to their smaller support, but whose size exceed three 32×32 blocks, it is more robust to adopt a localized motion estimation approach. Accordingly, a more simplified motion model that estimates only translation and divergence [35] is fitted to these regions.

By this stage, the only regions with no motion models are those that are neither large nor conformed to the dominant and major secondary motions. We assign fitting 2-D translation

motions onto the edge pixels of these unassigned regions at the 32×32 block level. Due to the occurrences of occlusion or multiple motions within a block, the motion computed from the block estimation step may be spurious or wrongly assigned. Thus each block is further tessellated into 8×8 "block-lets" to further refine motion assignment. The idea is to assign to the blocklet an optimal motion model from the motions of the 32×32 blocks neighboring the blocklet, as defined by a smoothness function.

Since optical flow is usually smooth, we can assume the true motion model of any blocklet i is within the candidate set $\mathbf{CS}(i)$ of motion models obtained from the nine neighboring 32×32 blocks. All the unassigned pixels—not only edge pixels—of each blocklet i are greedily initialized to the motion from $\mathbf{CS}(i)$ that maximizes the average pixel weight. Let $\mathbf{NBLM}(i)$ be the set of optimal motion models currently adopted by the eight neighboring blocklets of blocklet i , and $w_{blk}(a, m_b)$ the average pixel weight of blocklet a under motion m_b . Then during each of the seven iterations of the optimal motion selection process, each blocklet will in scan-line order select the motion j from $\mathbf{CS}(i)$ that best minimizes the following objective function over all j :

$$\text{OBJ}(i) = \zeta (w_{blk}(i, m_i) - w_{blk}(i, m_j)) + \sum_k^{\mathbf{NBLM}(i)} f_{\text{mtnDist}}(m_j, m_k) \quad (1)$$

$$f_{\text{mtnDist}}(m_j, m_k) = [1 \ 1 \ bs \ bs \ bs \ bs] |m_j - m_k|$$

where $\zeta = 30$, bs is the block size = 32, j indexes into $\mathbf{CS}(i)$, current motion $m_i = [a_{1,i} \ a_{2,i} \ a_{3,i} \ a_{4,i} \ a_{5,i} \ a_{6,i}]^T$, and candidate motion $m_j = [a_{1,j} \ a_{2,j} \ a_{3,j} \ a_{4,j} \ a_{5,j} \ a_{6,j}]^T$. The objective function comprises a smoothness term $f_{\text{mtnDist}}(m_j, m_k)$ that measures the dissimilarity between neighboring blocklet motion models and a data term to minimize the difference between the current block-let motion and the selected motion.

Finally, to select the best motion model for those regions which have been tessellated in blocks, we define the notion of a region motion consistency (RMC) measure, which computes the consistency of any particular motion model in a region. The idea is to implement a voting scheme where the motion most consistent with other models within the same region is chosen as the optimal region motion model. Let a region i contain N_m different motion models in the set $\mathbf{RM}(i)$ amongst its pixels, and q indexes $\mathbf{RM}(i)$. Let the membership of pixels belonging to the motion m_q have a cardinality of $|m_q|$. Then, the RMC(p) for motion model p is defined as

$$\begin{aligned} \text{RMC}(p) &= \sum_q^{N_m} \min(|m_p|, |m_q|) * \text{reg}_{\text{mtnDist}}(m_p, m_q) \\ &\text{reg}_{\text{mtnDiff}}(m_p, m_q) \\ &= \begin{cases} 0, & f_{\text{mtnDiff}}(m_p, m_q) \geq 2 \\ 1 - (f_{\text{mtnDiff}}(m_p, m_q)/2), & f_{\text{mtnDiff}}(m_p, m_q) < 2. \end{cases} \end{aligned} \quad (2)$$

The motion model p within $\mathbf{RM}(i)$ with the highest RMC is then adopted as the region motion model. Robust and dense optical flow has been obtained with these steps.

3) *Region Labeling with MRF*: We incorporate a few crucial cinematographic constraints previously alluded to into an MRF framework for motion segmentation. Firstly, we assume the FOA to be in the foreground. While there are some exceptions, it holds true for the vast majority of cases. Secondly, it is assumed that the dominant motion is sufficiently accurate to describe either the FOA or the background. Although the background usually conforms quite rigidly to one motion, this may not be the case for the FOA, which, for instance, may be a walking human with multiple articulated hand motions. For our shot indexing purpose, it is nevertheless sufficient to identify the dominant motion exhibited by the human body as foreground, while leaving out the swinging hands.

In accordance with these two constraints, we cast the region labeling problem as one where each region is labeled either as foreground or background. We also need to determine if the region's motion belongs to that of the "dominant motion" or "all other motions." Thus at the global level, the region labeling process has to test two hypotheses, *Hypo1* (dominant = *BG*, others = *FG*) and *Hypo2* (dominant = *FG*, others = *BG*). The hypothesis with the higher probability is taken to be the correct interpretation.

Applying MRF modeling [34], [36] to our region labeling problem, we represent the N regions $\mathbf{R} = \{R_1, R_2, R_3, \dots, R_N\}$ as the set of MRF sites, which are defined on a neighborhood system where physically adjacent sites are neighbors. A hypothesis of this MRF thus comprises the set of random variables, or configuration $\xi = \{\xi_1, \xi_2, \dots, \xi_N\}$, where ξ_i can take on either the *FG* or *BG* label for region i , as well as the variable H , which can take on *Hypo1* or *Hypo2*. Let the observation $\mathbf{O} = \{O_1, O_2, O_3, \dots, O_N\}$ be the set of individual features O_i observed for region R_i . The solution we seek is the optimal configuration ξ and hypothesis H that maximizes the MAP (maximum *a posteriori*) $P(\xi|\mathbf{O}, H)$, which can be expressed in the Bayesian framework [34], [36] as

$$P(\xi|\mathbf{O}, H) \propto \frac{e^{-U(\mathbf{O}, H|\xi)}}{Z_H} \quad (3)$$

where Z_H is the partition function and is constant for a specific H , while the MAP is maximized by minimizing the likelihood energy function $U(\mathbf{O}, H|\xi)$. $U(\mathbf{O}, H|\xi)$ comprises the intra-region and inter-region interactions captured by the local clique potentials V^D , V^S , and V^A to reflect our data, spatial, and attention constraints, respectively, detailed in the following sections. Let the set of all possible singleton cliques be $C1$ and pairwise cliques be $C2$ in this neighborhood system, then

$$\begin{aligned} U(\mathbf{O}, H|\xi) &= E_{\text{data}} + E_{\text{spatial}} + E_{\text{attention}} \\ &= \kappa_d \sum_{\{i\} \in C1} V_1^D(\xi_i, \mathbf{O}, H) \\ &\quad + \kappa_a \sum_{\{i\} \in C1} V_1^A(\xi_i, \mathbf{O}, H) \\ &\quad + \kappa_s \sum_{\{i, j\} \in C2} V_2^S(\xi_i, \xi_j, \mathbf{O}, H) \end{aligned} \quad (4)$$

where $\kappa_d = 5$, $\kappa_s = 2$, and $\kappa_a = 1.5$ are constants that control the relative importance of these three sets of energy

potentials. The clique potentials V_1^D and V_1^A are defined on singleton cliques, while V_2^S is defined on pairwise cliques, in accordance to the neighborhood system that satisfies the Markovianity condition for MRF representation [36].

a) *Data term*: The MRF solves two problems simultaneously: which regions conform to the dominant motion (the motion recovered first in the preceding section); and whether these regions belong to the foreground or background. In theory, the first task is readily determined by the $w(i)$ values computed for the dominant motion. $w(i)$ follows distinct likelihood distribution curves for both dominant motion regions (high likelihood) and nondominant motion regions (low likelihood), and the point where they intersect would be the optimal place to read off a decision threshold, which can be used for data energy term computation.

Empirically, this decision threshold varies strongly with two factors: global speed and the intensity gradient magnitude of each region. High global speed causes blurring of region boundaries, causing spuriously low $w(i)$. At the same time, low-intensity gradient magnitude for region i , $\text{avGrad}(i)$, leads to inaccurately high $w(i)$. To compensate for these factors, we compute an adaptive decision threshold $\text{EQ}_w(i)$ for each region i . For a frame with average dominant speed avGS , we define various speed levels at $\text{spd}_{\text{lo}} = 2$, $\text{spd}_{\text{mid}} = 4$, $\text{spd}_{\text{hi}} = 12$, and various intensity gradient magnitude levels at $\text{grad}_{\text{lo}} = 20$ and $\text{grad}_{\text{hi}} = 50$. Then we compute a global adjustment $\text{GS}_{\text{adjust}}$ that is adaptive to the various speed and gradient magnitude levels

$$\begin{aligned} \text{GS}_{\text{adjust}} &= \begin{cases} 0.1, & \text{avGS} \leq \text{spd}_{\text{lo}} \\ 0.1 * \frac{(\text{avGS} - \text{spd}_{\text{lo}})}{(\text{spd}_{\text{mid}} - \text{spd}_{\text{lo}})}, & \text{spd}_{\text{lo}} < \text{avGS} \leq \text{spd}_{\text{mid}} \\ -0.1 * \frac{(\text{avGS} - \text{spd}_{\text{mid}})}{(\text{spd}_{\text{hi}} - \text{spd}_{\text{mid}})}, & \text{spd}_{\text{mid}} < \text{avGS} \leq \text{spd}_{\text{hi}} \\ -0.1, & \text{avGS} \geq \text{spd}_{\text{hi}}. \end{cases} \end{aligned} \quad (5)$$

This global adjustment is finally added to $\text{EQ}_w(i)$ in a step that also compensates for possibly low $\text{avGrad}(i)$

$$\begin{aligned} \text{EQ}_w(i) &= \begin{cases} \text{EQ}_{wc} + \text{GS}_{\text{adjust}}, & \text{avGrad}(i) > \text{grad}_{\text{hi}} \\ \text{EQ}_{wc} + \text{GS}_{\text{adjust}} + \\ \quad (\text{avGrad}(i) - \text{grad}_{\text{lo}})/200, & \text{grad}_{\text{lo}} < \text{avGrad}(i) \leq \text{grad}_{\text{hi}} \\ \text{EQ}_{wc} + \text{GS}_{\text{adjust}} + 0.15, & \text{avGrad}(i) \leq \text{grad}_{\text{lo}} \end{cases} \end{aligned} \quad (6)$$

where $\text{EQ}_{wc} = 0.7$ is the equilibrium weight constant. After determining the suitable decision threshold $\text{EQ}_w(i)$ for every region i , the data energy potential is calculated as

$$\begin{aligned} V_1^D(\xi_i, \mathbf{O}, H) &= \begin{cases} -\frac{1}{2} \cdot \frac{w(i)}{\text{EQ}_w(i)}, & w(i) \leq \text{EQ}_w(i) \\ -\left(\frac{1}{2} + \frac{1}{2} \cdot \frac{w(i) - \text{EQ}_w(i)}{1 - \text{EQ}_w(i)}\right), & w(i) > \text{EQ}_w(i). \end{cases} \end{aligned} \quad (7)$$

In this context, $EQ_w(i)$ plays a very important role in the computation of a piecewise linear data energy function, as opposed to a binary function.

The heart of the occlusion handling mechanism, however, revolves around the computation of $w(i)$, where edge pixels are selectively summed depending on the hypothesized relative depth order of any region i with its adjacent regions. Edge pixels tend to be occluded whenever occlusion occurs, and as Bergen [37] observed, given the correct motions for any two regions, error density is often high on the occluded side of an edge, and low on the occluding side.

Thus, in any hypothesized set of segmentation labels, by excluding BG edge pixels that are adjacent to FG edge pixels when computing the data energy, occlusion becomes an integral part of the MRF process (Fig. 5) instead of an *ad hoc* or complicated multimotion occlusion/depth order handling process. Let $bdr_w(i, j)$ denote the total weights of the set of edge pixels along the borders of regions i and j , N_i the number of regions bordering i , and $av()$ be the average operator; then

$$\begin{aligned} w_{fg}(i) &= av \left(\sum_j^{N_i} bdr_w(i, j) \right) & \zeta_i &= FG \\ w_{bg}(i) &= av \left(\sum_j^{N_i} bdr_w(i, j) \cdot f_{occ}(i, j) \right) & \zeta_i &= BG \\ f_{occ}(i, j) &= \begin{cases} 0, & \zeta_i = BG, \zeta_j = FG \\ 1, & \text{other cases} \end{cases} \end{aligned} \quad (8)$$

Finally, the value of $w(i)$ is calculated differently depending on both the value H and the label ζ_i of region i using

$$w(i) = \begin{cases} w_{bg}(i), & H = Hypo1, \zeta_i = BG \\ 1 - w_{bg}(i), & H = Hypo1, \zeta_i = FG \\ 1 - w_{fg}(i), & H = Hypo2, \zeta_i = BG \\ w_{fg}(i), & H = Hypo2, \zeta_i = FG \end{cases} \quad (9)$$

b) Spatial term: The spatial energy potential, consisting of pairwise cliques, expresses the *a priori* assumption that adjacent regions with similar colors tend to share the same label (FG or BG) and encourage the formation of compact and adjoining foreground and background. The exact formulation of this spatial energy potential follows Tsai [34].

c) Attention term: As discussed in Section II, different classes of shot semantics direct viewer attention in their own characteristic manner. The “attention signature” of each shot can in turn be composed by correctly identifying the FOA in each frame and tracking the number of times it receives

attention throughout the shot. To model this attention process, we use two 2-D image buffers of the size of the image frame: Rec_{att} to record the net duration a pixel has been classified in the most recent 25 frames, subject to a ceiling of $T_{att_span} = 25$; and $Hist_{att}$ to record the total number of times the pixel has been classified as FG in the shot. The value of T_{att_span} for Rec_{att} is equivalent to approximately 1s in duration, and models the persistence behavior of attention span remaining on an area that has stopped moving before it fades.

Let $Rec_{att}(i)$ and $Hist_{att}(i)$ be the average value of the pixels of region i in the current Rec_{att} and $Hist_{att}$, respectively. All pixels in both buffers for each region are updated to their respective $Rec_{att}(i)$ and $Hist_{att}(i)$ values. Newly appearing pixels—those not mapped to by the optical flow from the previous frame—are excluded from computation.

Also, to model the phenomenon that recently moving objects are more likely to continue receiving attention, and to encourage smoothness in labeling along the temporal dimension, a threshold $T_{att} = 5$ is introduced which encourages FG labeling for a region with $Rec_{att}(i)$ above T_{att} while FG labeled regions if its $Rec_{att}(i)$ is below T_{att} . Note that T_{att} is set much lower than T_{att_span} to model the assumption that FOA attracts viewer’s attention faster than it is relinquished. To ensure stability, the attention energy potential is computed only after a burn-in period of the first T_{att_span} number of frame pairs. Expressing the above modeling assumptions, the attention energy potential term is

$$\begin{aligned} V_1^A(\zeta_i, \mathbf{O}, H) &= \begin{cases} -(\text{Rec}_{att}(i) - T_{att}) / \\ (T_{att_span} - T_{att}), & \zeta_i = FG \\ 0, & \zeta_i = BG \\ 0, & \text{if \#frames} < T_{att_span}. \end{cases} \end{aligned} \quad (10)$$

The configuration ζ is initialized with solitary clique (i.e., non-pairwise) data and attention energy potentials, and the ζ that minimizes $U(\mathbf{O}, H|\zeta)$ is found using the highest confidence first (HCF) method [38]. With the final region labelings, both attention buffers Rec_{att} and $Hist_{att}$ —denoted as AB in the equation below—are for the k th pixel as shown in (11) at the bottom of the page

Finally, the values in both buffers Rec_{att} and $Hist_{att}$ are shifted according to the motion model assigned to each pixel. To ameliorate the occasional spurious motion estimation and image segmentation, we perform memory smoothing on both Rec_{att} and $Hist_{att}$, using a 3×3 weighing kernel over pixels that belong to the same region. Fig. 6 shows some example attention signatures.

At the conclusion of the separate MAP maximization iterative processes for both $H = Hypo1$ and $H = Hypo2$,

$$\begin{aligned} AB_{att}[k] &= \begin{cases} AB_{att}[k] - \delta_{att} & \zeta[k] = BG \\ AB_{att}[k] + \delta_{att} & \zeta[k] = FG \end{cases} \\ \delta_{att} &= \begin{cases} 1, & |AB_{att}(i) - AB_{att}[k]| \leq T_{att} \\ 1 + \text{floor}(|AB_{att}(i) - AB_{att}[k]| / T_{att}), & |AB_{att}(i) - AB_{att}[k]| > T_{att} \end{cases} \end{aligned} \quad (11)$$

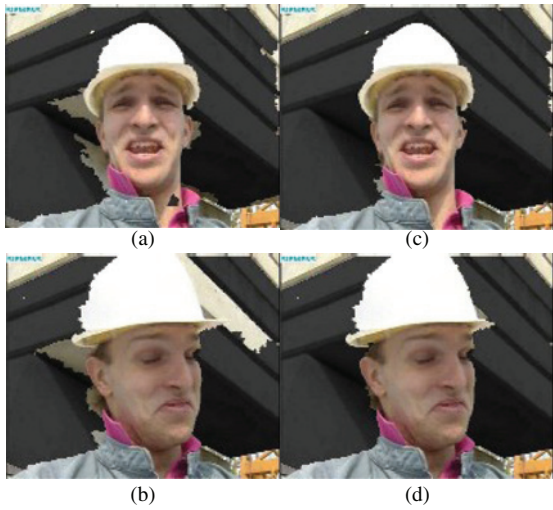


Fig. 5. Comparison of occlusion energy. Segmentation results with occlusion (c, d) and without occlusion (a, b) factored into the energy calculations.

the “weighted-likelihood” (WL) energy $U_{WL}(\mathbf{O}, H|\xi)$ is computed from the different final configurations ξ obtained under both H labels. $U_{WL}(\mathbf{O}, H|\xi)$ weighs the likelihood of each region i by its visual presence, as modeled by its number of edge pixels $\gamma(i)$. A WL partition function $Z_{WL,H}$ is computed under pseudo-likelihood assumptions [36]

$$\begin{aligned}
 &U_{WL}(\mathbf{O}, H|\xi) \\
 &= \sum_i^N \gamma(i) \left(\begin{array}{c} \kappa_d V_1^D(\xi_i, \mathbf{O}, H) + \kappa_a V_1^A(\xi_i, \mathbf{O}, H) \\ \kappa_s \sum_{j \in \text{Neigh}(i)} V_2^S(\xi_i, \xi_j, \mathbf{O}, H) \end{array} \right) \\
 &Z_{WL,H} = \exp \left(- \sum_{\xi_i \in \{BG, FG\}} U_{WL}(\mathbf{O}, H|\xi) \right).
 \end{aligned} \tag{12}$$

The WL MAP $P_{WL}(\xi|\mathbf{O}, H)$ is finally computed for both hypotheses, and the configuration ξ and hypothesis H responsible for the higher $P_{WL}(\xi|\mathbf{O}, H)$ are taken to be the truth labels

$$P_{WL}(\xi|\mathbf{O}, H) = \frac{e^{-U_{WL}(\mathbf{O}, H|\xi)}}{Z_{WL,H}}. \tag{13}$$

On a Pentium IV 3.4-GHz processor, the un-optimized C++ algorithm takes an average of 1.26s to compute dense optical flow and foreground/background segmentation for a 352×288 image.

C. Difficulties Encountered

Due to the extremely wide domain of Hollywood shots, there are four challenging circumstances where the region computed does not conform to the human perception of the FOA. Firstly, confusion between foreground and background regions typically occur when the background is extremely bland and small in area compared to the foreground. (e.g., close-up face shot framed by a bland wall). Without effective intensity gradient, the background tends to be assigned the wrong motion.

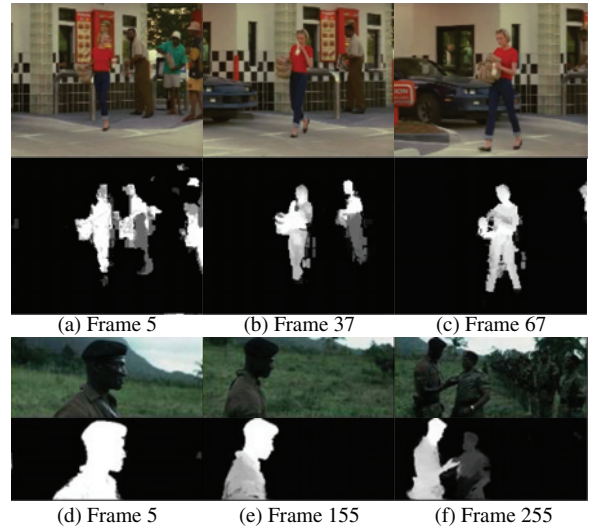


Fig. 6. Attention signature maps for sequences (a–c), (d–f). Whiter areas indicate higher attention intensity. Note the attention intensity rises and ebbs accurately according to the location of the FOA at the moment.

Secondly, occlusion of tracked FOA (i.e., a man who is tracked throughout the shot may be momentarily occluded with someone else walking in front of the camera) may wipe the accumulated memory away from the occluded foreground completely, rendering it useless to gauge the on-screen duration of the FOA. Thirdly, panoramic shots covering a wide range of depths and even indoor shots with cluttered scenes tend to result in severe violation of the affine motion model for the background. Finally, flow computation is also difficult in shots featuring non-Lambertian objects in motion (swirling water, burning fire, etc.), special lighting effects, or are simply too dim.

IV. FILM DIRECTING DESCRIPTORS

Having extracted the low level motion cues, the rest of this section details the various mid-level modules used to compute directing descriptors, from the outputs of the motion segmentation and region labeling modules (see the flowchart in Fig. 7). The various descriptors of a shot are finally concatenated into a 21-dimension shot descriptor vector (SDV) for final shot classification.

A. Key Frame and Frame Level Descriptors

Every shot is represented by a number of key frames. Starting with the first frame as a key frame, subsequent key frames are selected when the frame differencing threshold between the current frame under consideration and the immediate previous key frame exceeds a fixed threshold. Thus for every key frame, relative to the next key frame, we are able to obtain four separate types of motion-based information using the motion segmentation process described in the preceding section. These are the 1) dense optical flow, the 2) binary background/foreground image segmentation map, the 3) attention signature image map, and, finally, the 4) background motion, which is succinctly represented in the affine parametric form.

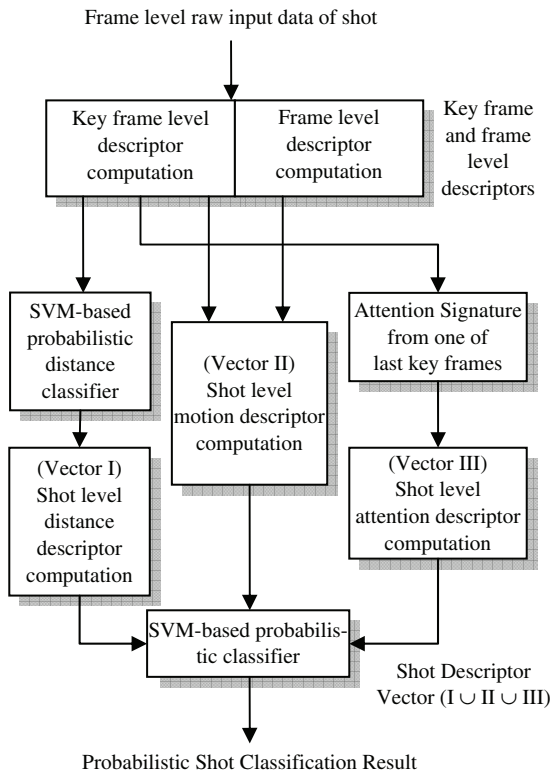


Fig. 7. Flowchart of the shot semantics classification process.

Since the raw motion-related information is extracted at the key frame level, it is necessary to interpolate and smooth the affine parameters and optical flow across the entire shot for every frame. Now, let every key frame be represented by a vector of descriptors, which we call the key frame descriptor vector (KFDV). These descriptors, which are computed from the above-mentioned key frame motion information, are as follows:

Background speed: The background speed, or the camera motion speed, for every key frame k in a shot is computed by $Mag_{BG,k} = \sqrt{a_1^2 + a_4^2}$.

Compressed foreground magnitude histogram (CFMH): In every key frame, our motion segmentation algorithm computes the optical flow within the foreground returned by the binary background/foreground image segmentation map. The velocities of the foreground pixels of every key frame are represented with a polar representation histogram comprising eight equi-angular bins of 45 degrees and 16 motion magnitude bins ([0.25, 1.0, 2.0, 3.0, 4.0, 5.0, 6.25, 7.75, 9.25, 11.0, 13.0, 15.0, 17.5, 20.5, 23.5, 30]), for a total of 128 bins. To obtain the final CFMH, the 128-bin histogram is merged into one bin along the angle dimension and merged along the magnitude dimension into only five bins of the following configuration: [1st–3rd bins, 4th–6th bins, 7th–9th bins, 10th–12th bins, and 13th–16th bins]. Higher foreground magnitudes have a loose correspondence with closer camera distances.

Motion vector entropy: This measure computes the entropy of the 128-bin version of the foreground polar representation histogram. To a certain extent, the entropy admits an indirect measure of the likelihood of disparate objects in the frame,

which in turn affects our inference of the number of objects, and hence the shot distance.

Foreground area percentage: This descriptor measures the total percentage of pixels designated as foreground w.r.t the frame area, and functions as the chief measure of camera distance. It is certainly true that the percentage of foreground is not strictly inversely proportional to camera shot distance. However, this descriptor functions adequately as a differentiator between the coarse camera distance categories of close-up/medium shots and long shots.

B. Shot Level Distance-Based Descriptor

In order to gauge the camera distance of every shot, the key frame level descriptors of the shot, or KFDV, are fed into a probabilistic SVM classifier (details in Section V-A). This SVM classifier is trained as a two-class classifier, using the close/medium and long shots as the two different classes. The percentages of key frames in a shot that are classified into the two classes are used as the shot level distance based descriptor. This descriptor is the chief means to distinguish between contextual-tracking and focus-tracking shots, which differ only in their camera distance.

C. Shot Level Motion-Based Descriptors

Normalized shot duration: This descriptor is derived by dividing the absolute duration of a shot (seconds) by $2t_{sd}$, where $t_{sd} = 4.3s$ is the average shot duration. Although semantics do not have any strict rules in film grammar concerning shot duration, chaotic shots tend to be of much shorter duration compared to most other semantic classes. On the other hand, both establishment and tracking shots tend to have a relatively long minimal duration, because both semantics require time to allow the viewers to be familiarized with the location or the tracked FOA respectively.

Stationarity percentage: This descriptor measures the percentage of frames in the shot where $Mag_{BG} < 2$. As opposed to other semantic classes, stationary shots tend to overwhelmingly cluster around high values of this measure.

Focus-in and focus-out percentages: The time-to-collision (TTC) value—time taken for an object to collide with camera given unchanged travel path—for every frame is computed as the well-known reciprocal of flow divergence given by $(1/(a_2 + a_6))$, and TTC values between [0, 400] and [-400, 0] are deemed significant indicators that the camera is experiencing the focus-in and focus-out phenomenon, respectively. Thus the percentages of frames within a shot deemed to undergo the focus-in and focus-out phenomena are the focus-in and focus-out percentages respectively. This is the main measure to differentiate between shots experiencing such phenomenon and those that are not.

Smoothness percentage: This descriptor calculates the longest consecutive period of the camera motion being either the pan a_1 or tilt a_4 without a change in direction, where the camera is deemed to be in motion if $Mag_{BG} > 2$. This descriptor serves as a measure of the smoothness of motion typical of establishment and tracking shots.

Shot CFMH: This descriptor is the average CFMH of every key frame within a shot. Histograms with very high probabilities residing in the high-magnitude bins usually belong to the chaotic semantic class, while the converse is true for static shots.

D. Shot Level Attention-Based Descriptor

Distinguishing between the establishment and tracking semantic classes, which share the same smooth background motion, requires the detection of the presence of an FOA. The attention image map, whose intensity at every pixel records the number of times it is classified as part of an FOA, directly addresses this challenge.

First, we normalize the values of the attention image map against the total number of key frames in a shot. This normalized map is used to construct an equally spaced 10-bin attention histogram $\mathbf{AH} = \{ah_1, ah_2, \dots, ah_{10}\}$ where each ah_n denotes the proportion of pixels in the attention image map with normalized attention values falling within the bin (e.g., ah_5 will have a bin range of [0.5, 0.6]). In theory, the attention histogram at the last key frame of the sequence should be able to give a good indication of FOA presence. However, there is a noticeable tendency for some tracking shots to allow the tracked FOA to either leave the frame or be occluded in the last moments. This would destroy the attention trail that had hitherto been maintained and give spurious classification results if only the last frame were used.

To counter this problem, we compute an \mathbf{AH} intensity measure \mathbf{AH}_{im} for each key frame in the last-third portion of the shot as $\mathbf{AH}_{im} = \sum_i^{10} (\text{median value of range of } ah_i) * ah_i$. For instance, bin ah_5 has the range [0.5, 0.6] and hence will have a median value of 0.55. The \mathbf{AH} with the highest \mathbf{AH}_{im} is finally used as the shot level attention-based descriptor. This ensures the FOA is at least consistently tracked until at least the last third of a shot to fulfill the tracking criterion, and can increase robustness against occlusions that occur in the last portion of the shot.

V. EXPERIMENTAL RESULTS

Our video corpus (Table II) comprises 5226 shots lasting 366 min and spans across seven movies of diverse genres: two full romantic comedy movies (*There's Something About Mary*, *Bedazzled*), one melodrama (*City of Angels*), and finally selected fast action shots from four action movies (*Lord Of the Rings*, *Star Wars*, *Golden Eye*, and *Starship Troopers*). From this video corpus, we have taken out 172 extremely low intensity shots whose average frame intensities are below 30, on the basis that these shots pose problems even for manual foreground and background segmentation. Other than this one condition, the video corpus has been chosen to maximize variety. Shot are manually segmented to avoid shot boundary errors. For labeling purposes, two persons were employed to independently label all shots according to guidelines in Section II. Finally, the few labeling discrepancies between both label sets are harmonized after discussion between the labelers. The number of shots with dual labels consists of 15.3% of the entire video corpus.

TABLE II
VIDEO CORPUS DESCRIPTION BY SHOT AND FRAMES

Movie	Shots (final/original)	Frames	Duration (min)
There's something about Mary	1004/1039	155938	104
Bedazzled	980/1012	119641	80
City of Angels	1053/1141	149545	100
Lord Of the Ring I	653/659	43211	29
Star Wars	495/502	27992	19
Golden Eye	565/568	37950	21
Starship Troopers	304/305	23171	13
Total	5054/5226	557448	366

TABLE III
COMPOSITION OF DIRECTING SEMANTIC CLASSES
IN VIDEO CORPUS (%)

	Static	F-Out	F-In	Estab	C-Track	F-Track	Chaotic
Number	1931	34	231	146	412	879	1421
(%)	38.21	0.67	4.57	2.89	8.15	17.39	28.12

A. Classification and Inference

Due to the highly irregular probability densities of the SDVs that represent each semantic category, the classifier used must be able to model extremely complicated class boundaries. We have therefore employed a specially adapted variant of the support vector machines (SVM), which can output a *posteriori* probabilities for multiple categories, permitting more refined classification.

To begin, SDVs are first normalized before being fed into two-class radial basis kernel SVMs for each of the 21 class pairs. In line with the ambiguous nature of those training data with dual labels, these data are included in the training set of both labels, and excluded only if the class pair coincides with the dual labels. Sigmoidals with adjustable parameters are then fitted to the decision values of the SVMs to learn and approximate accurate *a posteriori* distribution for each class pair [39].

An arriving test vector v_t is processed by each class-pair SVM to obtain the decision values, which are in turn fed into the respective sigmoidals to produce class-pair probabilities. These probabilities are finally combined together to compute the *a posteriori* of v_t for every category [40]. K-fold cross validation is used in a grid search to obtain the optimal penalty and margin parameters $C = 2$ and $\gamma = 5$.

However due to the great imbalance of samples between certain classes (Table III), it is necessary to use ensemble-based classification instead of straightforward SVM classification. For each training-classification iteration, we employ a shot selection method similar to bagging [41]. With this method, we randomly select from each class the smaller number between 300 and 85% of all shots of that class for training, and reserve the remaining shots for testing, allowing a much more balanced training set. Furthermore, for the larger training classes (>300 samples), which are not susceptible to the lack of samples, we have ensured that shots from the same movie are not selected into both the training and testing sets. During the testing phase, the test label of each test shot

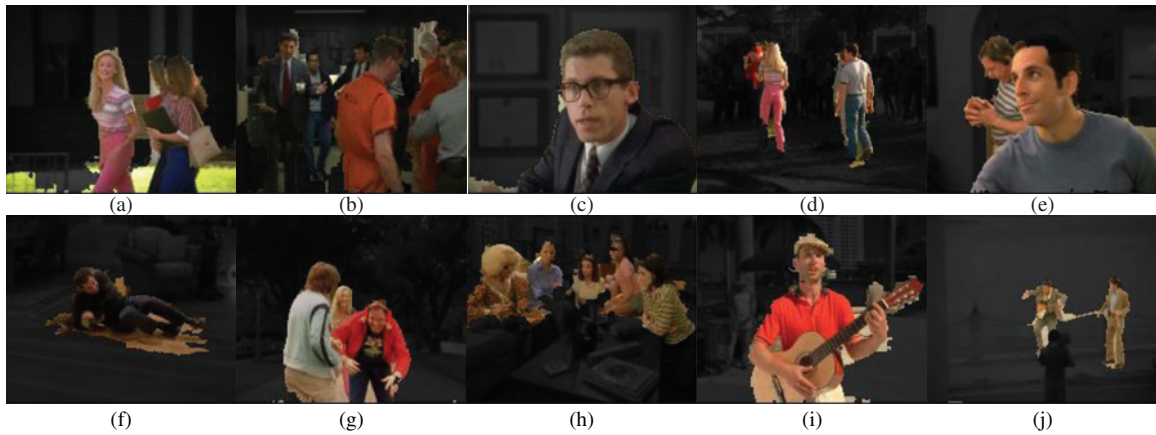


Fig. 8. Segmentation results from *There's Something About Mary*. Note the wide variety of camera distances present within these shots.

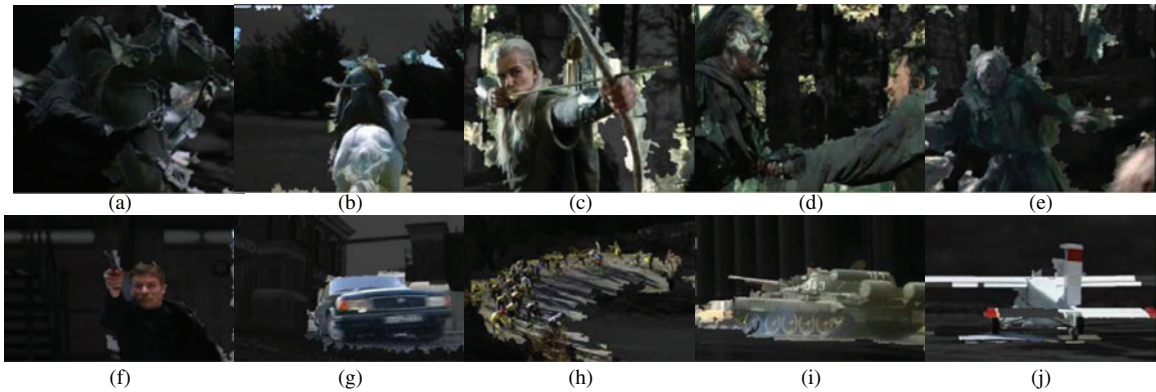


Fig. 9. Segmentation results from the action movies *Lord of the Rings: The Fellowship of the Ring* and *Golden Eye*. Presented are the segmentation results of some of the most furious action sequences in either of these movies.

TABLE IV
CONFUSION MATRIX FOR DIRECTING SEMANTIC CLASSES (%)

	Static	F-Out	F-In	Estab	C-Track	F-Track	Chaotic
Static	91.92	0.20	0.78	0.73	0.67	0.36	5.33
F-Out	2.94	82.35	0.00	1.47	5.88	2.94	4.41
F-In	0.43	0.22	87.88	0.43	0.65	2.16	8.23
Estab	2.74	0.00	2.05	82.88	4.79	4.11	3.42
C-Track	0.73	0.24	0.49	5.82	87.62	4.37	0.73
F-Track	0.57	0.34	2.39	1.14	5.01	86.46	4.10
Chaotic	6.12	0.28	1.76	0.91	1.34	4.22	85.36

is the class receiving the highest probability. This training–classification process is iterated 100 times, with the results aggregated over all iterations. Each shot is finally assigned to the test label receiving the most votes according to the aggregated results. The classification results are shown in the following tables.

B. Results

Some of the segmentation results from the experiments are illustrated in Figs. 8 and 9. Analyzing the confusion matrix in Table IV, certain classification error rates stand out and warrant an explanation. It is observed that the static and chaotic classes tend to be confused with one another. As the main

difference between these two classes lies in the magnitude of movement (see Section II-D), their labeling is open to certain amount of viewer subjectivity, and thus there will inevitably be a small number of “misclassified” borderline shots. However, such errors are not egregious in the sense that the amount of mistakes tends to fall within the margin of error expected in human interpretation over the semantic classes themselves.

Another source of relatively high error is that of establishment shots being mistaken as contextual-tracking (C-Track) shots. Both types of shots are characterized by long durations of panning motion and sometimes, if the small FOA moves too fast in a contextual-tracking shot, it is possible for the attention trail to vanish and consequently take on the appearance of an establishment shot, which does not have significant attention intensity.

Similarly, due to the similarities in the tracking motion, both contextual-tracking and focus-tracking classes have the tendency to be confused with one another. From the last column, it is noticed that chaotic class seems most prone to confusion with other classes. This is likely because chaotic class is the most unconstrained and unstructured class both in terms of motion characteristics and camera shot distance. It thus occupies a disproportionately large area in the descriptor space and this increases the likelihood of encroaching upon other classes.

From the first row of Table V, the recall rates for all classes seem satisfactory. However, due to the disproportionately

TABLE V
RECALL AND PRECISION FOR DIRECTING SEMANTIC CLASSES (%)

	Static	F-Out	F-In	Estab	C-Track	F-Track	Chaotic
Recall	91.92	82.35	87.88	82.88	87.62	86.46	85.36
Precision	94.62	69.14	75.46	65.94	80.67	88.68	87.87

TABLE VI
CONFUSION MATRIX FOR DIRECTING SEMANTIC CLASSES WITH NO
OCCLUSION HANDLING (%)

	Static	F-Out	F-In	Estab	C-Track	F-Track	Chaotic
Static	91.20	0.21	0.78	0.73	0.67	0.36	6.06
F-Out	2.94	82.35	0.00	1.47	5.88	2.94	4.41
F-In	0.43	0.22	87.45	0.43	0.65	2.16	8.66
Estab	2.74	0.00	2.05	82.19	6.16	4.11	2.74
C-Track	0.97	0.24	0.73	6.80	85.44	5.10	0.73
F Track	0.57	0.34	2.39	1.14	5.12	86.12	4.32
Chaotic	6.12	0.28	1.76	0.91	2.60	6.97	81.35

small sample sizes in the video corpus for the semantic classes focus-in, focus-out, and establishment, their precision rates are extremely susceptible to false positives from other much larger classes. These false positives, though few in number, are sufficient to significantly reduce precision rates of smaller classes.

To evaluate the effectiveness of the proposed occlusion handling mechanism by the MRF-based region labeling algorithm, the mechanism is “switched off” by adopting the hypothesis that the dominant motion is the background all the time. The new results of such a change are tabulated in Tables VI and VII and these can be compared with its counterpart results of Tables IV and V. It can be observed that although there are little differences for most results, there is a discernible drop in classification rates when occlusion handling is “switched off” for the C-Track, F-Track, and chaotic classes.

Occlusion handling is specifically formulated to identify foreground from background. Therefore it is expected to turn in better classification rates in comparison to algorithms which assume that the dominant motion is always the background, especially for classes with a higher proportion of close-up shots featuring dominant foreground. As a matter of fact, close-ups are very heavily concentrated in the static and chaotic classes; even the F-Track class is comprised mostly of medium shots. In the event of confusion between background and foreground, static shots by virtue of their stationarity are not likely to be mislabeled as other classes, as seen from the recall rates for the chaotic class (Table VII). However, chaotic shots are much more liable to be labeled as tracking shots, especially for the F-Track class (seen from its precision rate in Table VII), due to the relatively large motion characteristics they share. It can be concluded that the occlusion handling mechanism does improve semantic classification accuracy under certain circumstances.

In absolute terms, the improvement is small. However, the occlusion detection mechanism is only expected to improve the classification results for the small number of shots where the “dominant motion as background” assumption is violated.

TABLE VII
RECALL AND PRECISION FOR DIRECTING SEMANTIC CLASSES WITH NO
OCCLUSION HANDLING (%)

	Static	F-Out	F-In	Estab	C-Track	F-Track	Chaotic
Recall	91.20	82.35	87.45	82.19	85.44	86.12	81.35
Precision	94.53	69.14	75.10	64.34	76.61	84.49	86.30

Seen in the context of these minority shots, the improvement in percentage accuracy is significant indeed.

A possible problem posed by the dual-labeled shots is that their descriptor values are likely to take on the average values of the constituent shot segments. This may increase classification error if a third class shares the same descriptor space now occupied by the dual-labeled shots. While we content ourselves with using the shot as a conventional unit of analysis, a possible solution is to automatically analyze the shots frame by frame and then segment the shot itself into their constituent parts with coherent directing characteristics. However, frame-by-frame analysis is more unstable and time consuming compared to key frame analysis, and the segmenting of the shot into its constituent parts introduces its own instability. Nevertheless, this direction is worth exploring in future works.

VI. CONCLUSION

In this paper, we have formulated a coherent taxonomy of film directing semantics based on directing tasks routinely performed by the directors and characterized via two key film directing elements: camera motion/FOA behavior and camera distance. We proposed a novel edge-based MRF region labeling technique that can identify the FOA using integrated occlusion modeling, as well as track viewer attention throughout the shot. Our experiments have shown that the motion-based characteristics of the directing elements within a shot are sufficient to index its directing semantics well, despite the fact that the classes themselves correspond to relatively high level and complex semantics. This can lead to interesting applications for video content management and processing.

For future works, it will be useful to investigate how to approximate camera distances reliably enough to tell apart close-ups from medium shots. For instance, the use of face detectors can conceivably give useful clues to the camera distance. Extending the idea further, contextual information such as inter-shot relationships can be incorporated into the framework, paving the way for an even richer semantics taxonomy. Other promising avenues for further research include the use of other modalities such as audio to extend the variety of semantics output.

REFERENCES

- [1] *The MPEG-7 visual part of the XM 4.0*, HI, ISO/IEC MPEG99/W3068, 1999.
- [2] C. G. M. Snoek and M. Worring, “Multimodal video indexing: A review of the state-of-the-art,” *Multimedia Tools Applicat.*, vol. 25, no. 1, pp. 5–35, Jan. 2005.
- [3] M. Lazarescu and S. Venkatesh, “Using camera motion to identify types of American football plays,” in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 2, Jul. 2003, pp. 181–184.

- [4] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga, "Sports video categorizing method using camera motion parameters," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 2, Jul. 2003, pp. 461–464.
- [5] B. T. Truong, S. Venkatesh, and C. Dorai, "Discovering semantics from visualizations of film takes," in *Proc. IEEE Int. Conf. Multimedia Model.*, Jan. 2004, pp. 109–116.
- [6] F. Nack and A. Parkes, "The application of video semantics and theme representation in automated video editing," *Multimedia Tools Appl.*, vol. 4, no. 1, pp. 57–83, Jan. 1997.
- [7] R. Hamoumd and R. Mohr, "Interactive tools for constructing and browsing structures for movie films," in *Proc. 8th ACM Int. Conf. Multimedia*, Marina del Rey, CA, 2000, pp. 497–498.
- [8] S. Dagtas, W. Al-Khatib, A. Ghafoor, and R. L. Kashyap, "Models for motion-based video indexing and retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 88–101, Jan. 2000.
- [9] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.
- [10] F. M. Idris and S. Panchanathan, "Spatio-temporal indexing of vector quantized video sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 5, pp. 728–740, Oct. 1997.
- [11] J. H. Oh, M. Thenneru, and N. Jiang, "Hierarchical video indexing based on changes of camera and object motions," in *Proc. ACM Symp. Appl. Comput.*, Melbourne, FL, 2003, pp. 917–921.
- [12] R. Fablet, P. Boutheymy, and P. Perez, "Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 393–407, Apr. 2002.
- [13] P. Over, T. Ianeva, W. Kraaijz, and A. F. Smeaton, "TRECVID 2005—An overview," in *Proc. TRECVID 2005*, 2006.
- [14] Y.-H. Ho, C.-W. Lin, J.-F. Chen, and H.-Y. M. Liao, "Fast coarse-to-fine video retrieval using shot-level spatio-temporal statistics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 642–648, May 2006.
- [15] V. Mezaris, I. Kompatsiaris, N. V. Boulgouris, and M. G. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 606–621, May 2004.
- [16] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and S. Zhong, "VideoQ: An automated content based video search system using visual cues," in *Proc. 5th ACM Int. Conf. Multimedia*, Seattle, WA: ACM, 1997, pp. 313–324.
- [17] Y.-F. Ma and H.-J. Zhang, "Motion pattern based video classification using support vector machines," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, Phoenix-Scottsdale, AZ, May 2002, pp. 69–72.
- [18] C. Dorai and S. Venkatesh, *Media Computing-Computational Media Aesthetics*. 1st ed. Norwell, MA: Kluwer, 2002.
- [19] D. Bordwell and K. Thompson, *Film Art: An Introduction*. 7th ed. New York: McGraw-Hill, 2004.
- [20] S. D. Katz, *Film Directing Shot by Shot: Visualizing from Concept to Screen*. Studio City, CA: Michael Wiese Productions, 1991.
- [21] J. Monaco, *How to Read a Film: Movies, Media, Multimedia*. 3rd ed. London, U.K.: Oxford Univ. Press, 2000.
- [22] D. Arijon, *Grammar of the Film Language*. Los Angeles, CA: Silman-James Press, 1976.
- [23] J.-M. Odobez and P. Boutheymy, "Robust multiresolution estimation of parametric motion models," *J. Visual Commun. Image Representation*, vol. 6, no. 4, pp. 348–365, Dec. 1995.
- [24] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 625–637, Sep. 1994.
- [25] H. Xu, A. A. Younis, and M. R. Kabuka, "Automatic moving object extraction for content-based applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 796–812, Jun. 2004.
- [26] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Video object segmentation using bayes-based temporal tracking and trajectory-based region merging," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 782–795, Jun. 2004.
- [27] D. S. Tweed and A. D. Calway, "Integrated segmentation and depth ordering of motion layers in image sequences," *Image Vision Comput.*, vol. 20, no. 9–10, pp. 709–723, 2002.
- [28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
- [29] A.-R. Mansouri and J. Konrad, "Multiple motion segmentation with level sets," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 201–220, Feb. 2003.
- [30] H. Sekkati and A. Mitiche, "Joint optical flow estimation, segmentation, and 3d interpretation with level sets," *Comput. Vision Image Understand.*, vol. 103, no. 2, pp. 89–100, Aug. 2006.
- [31] P. Smith, T. Drummond, and R. Cipolla, "Layered motion segmentation and depth ordering by tracking edges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 4, pp. 479–494, Apr. 2004.
- [32] I. Patras, E. A. Hendriks, and R. L. Lagendijk, "Video segmentation by MAP labeling of watershed segments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 326–332, Mar. 2001.
- [33] P. H. S. Torr, R. Szeliski, and P. Anandan, "An integrated Bayesian approach to layer extraction from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 297–303, Mar. 2001.
- [34] Y. Tsaig and A. Averbuch, "Automatic segmentation of moving objects in video sequences: A region labeling approach," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 12, no. 7, pp. 597–612, Jul. 2002.
- [35] P. De Smet and D. De Vleeschauwer, "Performance and scalability of a highly optimized rainfalling watershed algorithm," in *Proc. Int. Conf. Imaging Sci., Syst. Technol. (CISST)*, 1998, pp. 266–273.
- [36] Z. Li Stan, *Markov Random Field Modeling in Image Analysis*. 1st ed. New York: Springer-Verlag, 1995.
- [37] L. Bergen and F. Meyer, "A novel approach to depth ordering in monocular image sequences," in *Proc. IEEE Intl. Conf. Comput. Vision Pattern Recognition*, vol. 2, Hilton Head Island, SC, Jun. 2000, pp. 536–541.
- [38] P. B. Chou and C. M. Brown, "The theory and practice of Bayesian image labeling," *Intl. J. Comput. Vision*, vol. 4, no. 3, pp. 185–210, 1990.
- [39] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Ann. Statist.*, vol. 26, no. 2, pp. 451–471, 1998.
- [40] J. C. Platt, *Advances in Large Margin Classifiers*, 1st ed. Cambridge, MA: MIT Press, 1999, pp. 61–74.
- [41] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.



Hee Lin Wang received the B.Eng. degree and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2001 and 2008, respectively.

He is currently a research fellow at the Institute of Infocomm Research, Agency for Science, Technology and Research, Singapore. His research interests include fingerprint biometrics, affective classification, multimedia indexing, Markov Random Field augmented particle-filter tracking, augmented reality, adaboost classification, person categorization using face and motion

segmentation.



Loong-Fah Cheong was born in Singapore on June 1, 1965. He received the B.Eng. degree from the National University of Singapore, Singapore, and the Ph.D. degree in computer science from the Center for Automation Research, University of Maryland, College Park, MD, in 1990 and 1996, respectively.

In 1996, he joined the Department of Electrical and Computer Engineering, National University of Singapore, where he is currently an Associate Professor. His research interests are related to the basic processes in the perception of 3-D motion, shape and their relationship, as well as the application of these theoretical findings to specific problems in navigation and in multimedia systems, for instance, in the problems of video indexing in large databases.