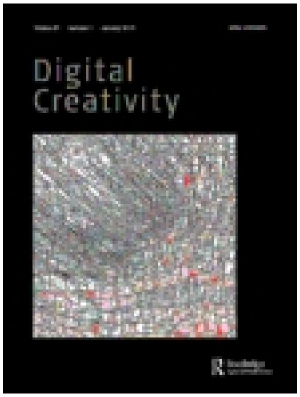

Digital Creativity
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ndcr20

Retrieval of motion composition in film

Matthias Zeppelzauer [a], Maia Zaharieva [a], Dalibor Mitrović [a] & Christian Breiteneder [a]

[a] Interactive Media Systems Group, Vienna University of Technology

Available online: 04 Jan 2012

To cite this article: Matthias Zeppelzauer, Maia Zaharieva, Dalibor Mitrović & Christian Breiteneder (2011): Retrieval of motion composition in film, Digital Creativity, 22:4, 219-234

To link to this article: http://dx.doi.org/10.1080/14626268.2011.622282

Digital Creativity
2011, Vol. 22, No. 4, pp. 219–234

Retrieval of motion composition in film

Matthias Zeppelzauer, Maia Zaharieva, Dalibor Mitrović and Christian Breiteneder

Interactive Media Systems Group, Vienna University of Technology

zeppelzauer@ims.tuwien.ac.at; zaharieva@ims.tuwien.ac.at; mitrovic@ims.tuwien.ac.at; breiteneder@ims.tuwien.ac.at

Abstract

This article presents methods for the automatic retrieval of motion and motion compositions in movies. We introduce a user-friendly sketch-based query interface that enables the user to describe desired motion compositions. Based on this abstract description, a tolerant matching scheme extracts shots from a movie with a similar composition. We investigate and evaluate two application scenarios: the retrieval of motion compositions and the retrieval of matching motions (a technique in continuity editing). Experiments show that the developed methods accurately and promptly retrieve relevant shots. The presented methods enable new ways of searching and investigating movies.

Keywords: motion retrieval, query-by-motion, motion composition, matching action, motion segmentation

1 Introduction

Motion characterises the style of a movie and significantly influences the way it is perceived by viewers. Motion controls tension and tempo in a scene, creates visual rhythm and gives a movie temporal continuity. Furthermore, motion gives evidence about the genre of a movie and about the director's style.

Motion is of special interest to film-makers and film theorists—however from different perspectives. Film-makers employ motion to increase the tension in a scene, for example. This may be achieved by using a shaky camera (steadicam) and showing fast-moving objects. Film theorists otherwise reverse the process of film-making and manually analyse in detail; for example, the usage of camera and object motion in order to investigate the progression of tension over an entire movie.

Automatic methods for motion retrieval support both film theorists and film-makers in their creative work by allowing efficient search and retrieval of particular types of motions and motion compositions. In particular, film theorists manually analyse motion shot by shot and in great detail, which is a tedious, time-consuming and error-prone task. Automatic motion retrieval supports the expert in finding typical motion directions, locations and combinations of interest more efficiently and sometimes also more accurately. Film-makers benefit from motion retrieval as it supports editors and directors in searching for alternative ways of motion editing. Motion

ISSN 1462-6268 © 2011 Taylor & Francis
http://dx.doi.org/10.1080/14626268.2011.622282
http://www.tandfonline.com

retrieval may further enhance the post-production of movies by automatically checking the quality of motion continuity, detecting errors and jumps. Additionally, motion analysis and retrieval may be incorporated into production tools which automatically generate movies from large movie and video databases.

Different approaches for the automatic analysis and retrieval of motion have been proposed in literature. However, they are hardly suitable for the applications of film theorists and filmmakers for different reasons. Most methods are based on the analysis of object motions only. They first segment and track the objects in a shot (Chang *et al.* 1998, Buzan *et al.* 2004) and then compute a motion trajectory for each object (Dagtas *et al.* 2000). Retrieval is then performed by matching the object trajectories with trajectories provided by the user (Bashir *et al.* 2007). However, some types of motion which are important for motion compositions cannot be represented by these approaches because they are difficult to segment, such as groups of objects (e.g. people, cars), motion of water (e.g. rivers) and smoke. Other approaches skip object segmentation and focus on the retrieval of camera motions only (Ardizzone *et al.* 1996, Hanis and Sziranyi 2003). Such methods rely on quantitative representations of motion which are too coarse and inaccurate for the retrieval of motion compositions.

This article introduces a more general approach to motion retrieval. The presented methods build upon a technique for motion segmentation that we originally presented in 2010 (Zeppelzauer *et al.* 2010). The technique clusters previously tracked motion trajectories into meaningful motion segments. Motion segments represent object motions as well as camera motions and are a convenient basis for the retrieval of motion. We define a novel type of query for the description of user-defined motion compositions. Based on the query, a tolerant matching scheme extracts the shots in a movie which are most similar. The method enables searching for typical camera motions, object motions and characteristic motion directions. Finally, we extend the method to the retrieval of *matching motion* (see Section 3.3) by adapting the query model and refining the matching scheme.

We demonstrate the performance of motion composition retrieval in the context of historical documentaries, since they contain numerous (repeated) complex compositions. The retrieval of matching motion is demonstrated with the contemporary movie *Run Lola Run* by Tom Tykwer from 1998. *Run Lola Run* is a thriller where the director makes extensive use of matching on action throughout the film to create the impression of a seamless storyline.

2 Motion analysis

A movie consists of different hierarchical levels. The basic visual unit in a movie is the *frame*—a single still image on the strip of film. Several frames make up a *shot*. In the phase of shooting, a shot is ‘one uninterrupted run of the camera to expose a series of frames. Also called a *take*’ (Bordwell and Thompson 2008, p. 8). Later, in the edited film a shot is ‘the length of film from one splice or optical transition to the next’ (Beaver 2009, p. 230). The latter definition is relevant for us, since we analyse already edited movies. Two shots (after editing) are connected either by an abrupt transition (*cut*) or by a gradual transition (e.g. dissolve and fade). The next higher structural level is the *scene*. A scene is ‘a unit of a motion picture usually composed of a number of interrelated shots that are unified by location or dramatic incident’ (Beaver 2009, p. 223).

We first temporally partition the movie into its shots by the method proposed in Zeppelzauer *et al.* (2008). For each shot motion tracking and motion segmentation is then performed separately.

2.1 Motion tracking

Different methods exist for the estimation and tracking of motion. The basis for motion segmentation is usually a dense motion field obtained from an optical flow estimator (e.g. Horn and Schunck 1981). Optical flow methods estimate the motion between two successive frames at

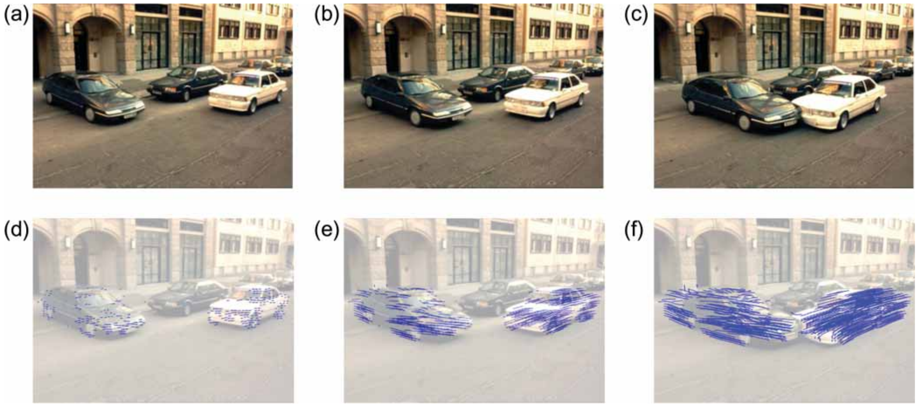

Figure 1. Motion trajectories obtained from the KLT feature tracker.
Credit: Stadtkino Filmverleih; mark-up: Matthias Zeppelzauer. Copyright by Stadtkino Filmverleih.

each pixel. This results in a dense motion field for each frame. However, high-quality optical flow estimators are too time-consuming to be applied to a full-length feature film (Brox *et al.* 2004).

Feature trackers are more suitable for this task since they can operate in real-time. Feature trackers first select *distinct* points in a frame (e.g. corners of objects) and then attempt to trace these points over time in subsequent frames. Points that cannot be tracked any further (e.g. because they move out of the frame or get occluded) are replaced continuously by new points. Figure 1 shows the trajectories of tracked feature points for an example sequence where two cars move towards each other and finally crash.

Feature tracking generates trajectories that represent the motion performed by the selected feature points. In contrast to optical flow, the resulting motion field is sparse (trajectories exist only for a few points and not for all pixels). We employ the Kanade–Lucas–Tomasi feature tracker (KLT) because of its efficiency and its ability to robustly track feature points across multiple frames (Shi and Tomasi 1994).

2.2 Motion segmentation

Motion segmentation aims at grouping motion trajectories that originate from the same source (an object or a camera movement). This segmentation allows an abstract description of the motion content in a shot and can be used for the retrieval of typical motions and motion compositions (see Section 3).

Segmentation of sparse motion fields obtained from feature tracking is more complex than of dense fields. The trajectories are different in length and have varying begin and end times. Due to occlusions and tracking errors, the trajectories often break off. Conventional segmentation methods cannot be applied to such heterogeneous data because they cannot cope with missing and incomplete data. This motivated the development of a novel method for the segmentation of fragmented motion fields (Zeppelzauer *et al.* 2010). The idea is to segment the entire sparse field of trajectories directly by iteratively grouping temporally overlapping trajectories. The process is illustrated in Figure 2. There are two stages. The first stage (*iterative clustering*) groups temporally overlapping trajectories with similar velocity direction and magnitude. The second stage (*merging*) connects the segments from the first stage into temporally adjacent segments covering larger time spans.

A basic assumption is that trajectories that perform similar motion at the same time belong to the same motion segment. According to this definition, a segment can represent motion of a single object, a group of several similarly moving objects, as well as motions of the camera. Conse-

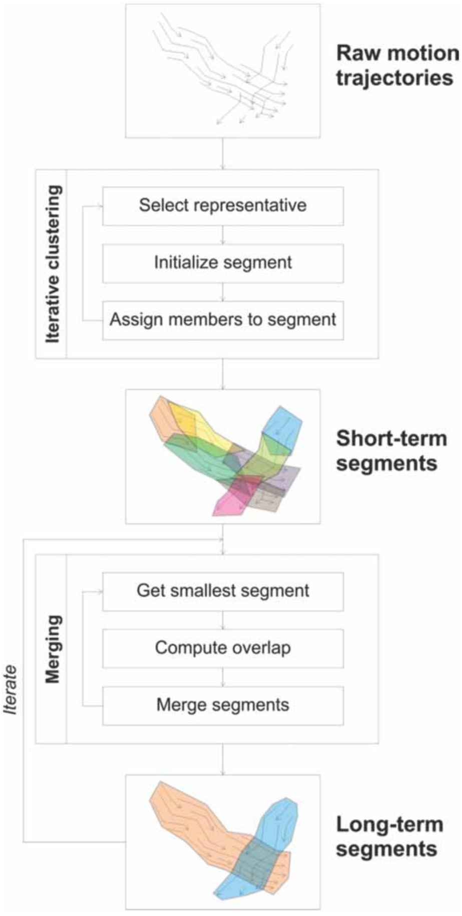

Figure 2. The process of motion segmentation.

quently, segmentation is not restricted to a particular type and source of motion which is important for the retrieval of arbitrary motion compositions.

The first stage of the approach (iterative clustering) starts with the selection of expressive trajectories (representatives) for the initialisation of new motion segments. A good representative is a trajectory that travels a significant distance. In the next step, all temporally overlapping trajectories which are similar (in terms of velocity direction and magnitude) to the selected representative are added to the new segment. Finally, all trajectories that lie temporally fully inside the new segment are removed from the original motion field. Trajectories that are temporally not fully covered by the segment *remain* in the motion field. That enables trajectories to be assigned to multiple temporally adjacent segments in further iterations. This is an important prerequisite for the creation of long-term segments in the second stage of the algorithm.

The next iteration starts with the selection of a new representative trajectory from the motion field. The algorithm terminates when no more trajectories are left in the motion field. Iterative clustering generates a set of *overlapping* motion segments. The temporal extent of the segments tends to be rather short (it is limited by the temporal extent of the representative trajectories). Consequently, the iterative clustering yields an over-segmentation of the motion field, as illustrated in Figure 2.

The goal of *merging* is to reduce the over-segmentation by connecting temporally adjacent segments that represent the same motion. This is performed by hierarchically merging overlapping segments. Beginning with the smallest segment, we search for the one which shares the most trajectories with this segment. If the portion of shared trajectories exceeds a certain threshold they are merged. In the following, this process is repeated successively with the next larger segment until all segments have been considered for merging. The result is a smaller set of segments with larger temporal extent. Further iterations (see the arrow labelled '*iterate*' in Figure 2) repeatedly merge newly created segments until no segments can be merged any more. The ultimate result of the merging procedure is a small set of segments which represent long-term motion.

An example for the segmentation of different object motions is illustrated in Figure 3. The shot shows a group of people walking up a hill. The people in the group first move away from the camera, towards the hill (in the lower right quarter), then turn to the left, walk up the hill and finally vanish behind the hill. At the end of the shot a horse enters at the top of the hill in the opposite direction (annotated with the short arrow in Figure 3(c)). From the three keyframes 3(a)–3(c) we observe a number of artefacts (flicker, scratches and dirt). Since motion tracking is sensitive to intensity variations like flicker, the

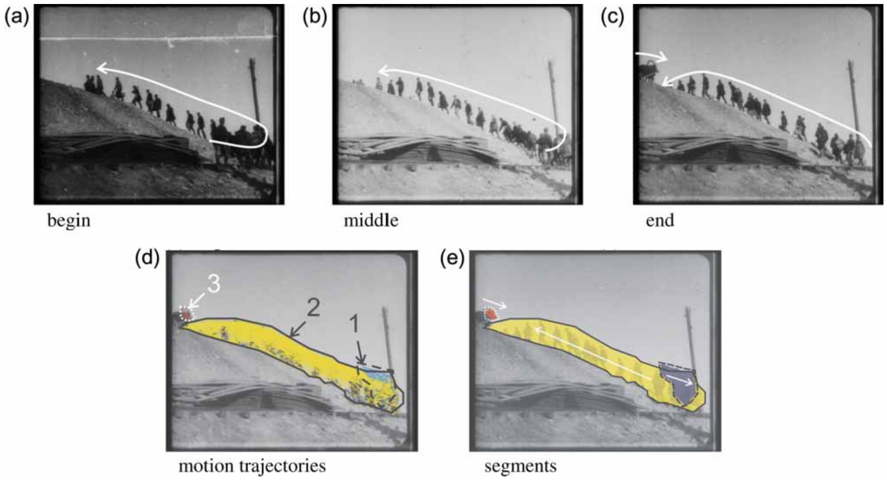

Figure 3. Motion segmentation results. Images 3(a)–3(c) are keyframes from the beginning, middle and end of the shot (white arrows represent the motion directions). Figure 3(d) shows the clustered trajectories and image 3(e) illustrates the three resulting motion segments (arrows indicate the primary motion direction of a segment).
Source: Austrian Film Museum; mark-up: Matthias Zeppelzauer. Copyright Austrian Film Museum.

trajectories are noisy and frequently break off. However, the approach is able to create temporally coherent motion segments. The movement of the group of people away from the camera is represented by segment 1 in Figure 3(d). Segment 2 represents the motion of the people walking up the hill. The third segment represents the horse that enters at the end of the shot from the left. The result shows that the approach is able to segment large groups of objects as well as small individual objects.

3 Motion retrieval

The obtained motion segments allow for a compact and expressive description of the motion contained in a shot. The segments describe diverse types of motion and enable the development of novel applications for retrieval and analysis of motion in a movie.

We investigate two different application scenarios that build upon the previously extracted motion segments. In the first scenario the goal is the retrieval of shots with a particular *motion composition* consisting of camera motions, object motions or both (see Section 3.2). The second scenario targets the retrieval of pairs of shots that are connected by *matching action* (see Section 3.3). In both application scenarios the user has to define a query. The design of the query is a crucial factor since it defines in which way the user has to specify the motion content of interest. Different types of queries exist in retrieval, such as textual queries, example-based queries and sketch-based queries. In the following, we define an intuitive type of sketch-based query that enables the user to specify arbitrary search requests.

3.1 Query formulation

A number of systems incorporating motion for retrieval have been introduced, for example *VideoQ* (Chang *et al.* 1998), *MovEase* (Ahanger *et al.* 1995), *Picturesque* (Dagtas *et al.* 2000) and the system described by Dimitrova and Golshani (1995). These systems require the user to define trajectories that represent the motion of the objects contained in the sequence of interest. Trajectory-based motion queries can become very complex with an increasing number of available degrees of freedom (Ahanger *et al.* 1995). Each moving object may be described, for example, by its trajectory together with its projected size, direction, speed and acceleration at each time instant. The definition of such a query can become counter-intuitive and time-consuming for complex motion compositions. Tolerant retrieval is difficult due to the large amount of detail contained in the motion description. Furthermore, the definition of such a motion query requires detailed knowledge about the sequences of interest, which reduces the exploratory capabilities of the corresponding systems.

We envision a query that is much easier (and faster) to define and that allows the user to integrate more variability into the motion description. We perform experiments with six test users who are all film experts. They are asked to sketch the motion content of selected shots on a piece of paper. The resulting sketches reveal that the most important information for the participants is the direction of the motion followed by its spatial location. Velocity and acceleration are neglected by most users. Even the size of the moving objects plays a secondary role. Furthermore, most test persons sketch motions of groups of objects as one coherent motion.

Based on these findings, we develop a query that enables sketching the motions of interest as vectors in a sketch-pad window. The absolute position and the length of a vector coarsely specify the region where the corresponding motion occurs (see Figure 5 for example queries). For simplicity, we do not consider velocity magnitude and acceleration in the queries. Large moving objects or groups of objects can be specified by drawing several nearby vectors with similar direction. For spatially distributed motion like camera motion, the user simply sketches several arrows with the desired direction(s) distributed over the entire query window.[1] The proposed query allows for expressive motion descriptions and at the same time provides enough freedom for tolerant retrieval.

The numerical description of a query contains the directions of all provided query vectors. Additionally, a region (aligned rectangle or ellipse) around each vector is extracted that represents the area covered by the corresponding motion. These two parameters (per query vector) are sufficient to represent motion compositions.

3.2 Retrieval of motion compositions

The retrieval of motion compositions requires the matching of the query with the previously computed motion segments. Therefore, we extract representative information from the motion segments that is structurally similar to the parameters derived from the query vectors. For each motion segment two parameters are computed: a representative motion direction and the spatial region covered by the segment.

The segments obtained from motion segmentation comprise a sparse set of fragmented trajectories. The extraction of representative information for such a segment is difficult, since the trajectories have different begin and end times and are spatially distributed. We extract the *median direction* of all trajectories, which is a robust estimate for the dominant motion direction of the segment. The second extracted parameter is the region covered by a segment. Therefore, we compute the polygon that encompasses all trajectories of the segment.

In the next step, a match between the query and the motion segments is established based on the extracted directional and spatial information. Optionally, temporal parameters (start time and duration of a motion) can easily be incorporated as additional constraints. During matching, all motion segments of a shot are compared with the query. A query Q is defined as a set of N query vectors q_i with directions θ_i and assigned regions R_i. A shot S contains M motion segments m_j with representative (median) directions φ_j and regions (surrounding polygons) M_j. Matching between a query and the motion segments of a shot is performed in three stages.

In the first stage a matching score $s_{i,j}$ between each query vector q_i and each motion segment m_j of the shot is computed as:

$$s_{i,j} = \frac{\theta_i \cdot \varphi_j}{\parallel \theta_i \parallel \cdot \parallel \varphi_j \parallel} \cdot \frac{|R_i \cap M_j|}{|R_i|} \cdot \left(1 - \frac{|M_j \backslash R_i|}{|M_j|}\right) \quad (1)$$

The first term is the cosine similarity of the query vector's direction θ_i and the motion segment's median direction φ_j. The second term is the portion of intersection between the region covered by the query vector R_i and the motion segment's region M_j. The spatial intersection is negatively weighted (penalised) by the area of the motion segment *not* covered by the query vector: $|M_j \backslash R_i|$. Matching each query vector with each motion segment yields a set of scores $s_{i,1\ldots M}$ for each query vector q_i. The scores are in the range $[-1; 1]$, where a score of 1 signifies a perfect match and a score below 0 represents a negative match.

In the second stage, all positive scores for a query vector q_i are summed: $s_i = \sum_{j=1}^{M} \max(s_{i,j,} 0)$. This allows one query vector to score on several motion segments. Negative scores obtained due to negative cosine similarity are ignored. Finally, in the third stage an overall score s is obtained by taking the sum of the scores s_i of all query vectors. The shots with the highest overall scores for the query are returned to the user.

The matching procedure is tolerant, since it does not require *all* query vectors to match with a retrieved shot. Additionally, a query vector accumulates scores from all matching motion segments, which makes matching more robust under noisy conditions, where motions are sometimes split into multiple motion segments. The amount of desired tolerance depends on the application. Matching can be performed more strictly by introducing penalties for non-matched and poorly matched query vectors.

3.2.1 Evaluation

For the evaluation of the proposed method we employ archive film material. The films are historical artistic documentaries from the late 1920s. The films exhibit twofold challenges that originate from their technical and from their artistic nature. From the technical point of view, the film material contains numerous artefacts such as flicker, shaking and dirt that impede the process of motion tracking and segmentation. From an artistic point of view, the applied documentary technique is highly innovative and employs complex motion compositions; for example, hammering, camera travellings and contrapuntal movements as shown in Figure 4. See Zeppelzauer *et al.* (2008) for a detailed description of the material.

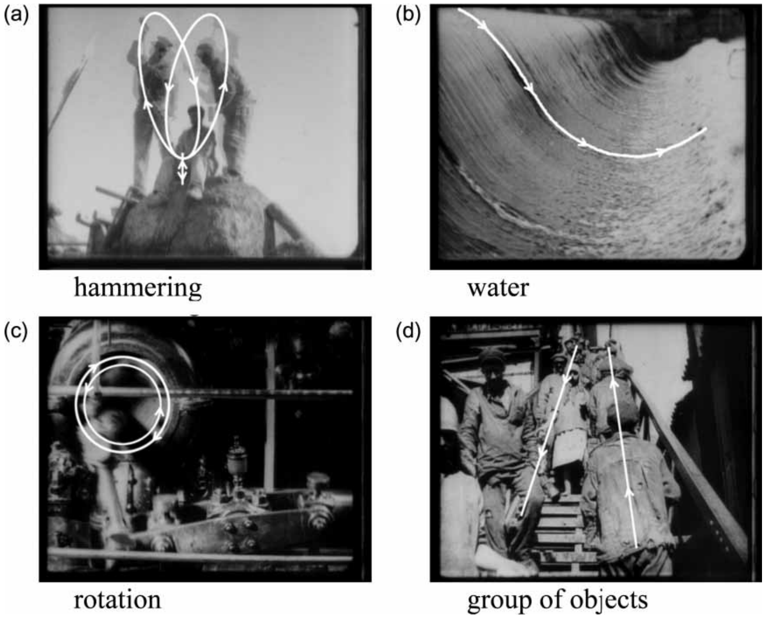

Figure 4. Typical motion compositions.
Source: Austrian Film Museum; mark-up: Maia Zaharieva. Copyright Austrian Film Museum.

We test the proposed method with queries representing camera motions, object motions, motions of groups of objects and combinations thereof. Experiments show that the method performs well for object motions. However, for the retrieval of spatially distributed motion (e.g. camera motions) which is represented by several query vectors the ranking of the retrieved shots is suboptimal. That means that the most relevant shots are indeed retrieved but do not yield the highest scores. This can be improved by incorporating the total number of scoring query vectors into the matching function from Equation 1. We weight the score s with the number of scoring query vectors $P_q : s_{total} = P_q \cdot s$. This increases the score of matches with a larger number of scoring query vectors and generates rankings that better correspond with the user's expectations.

We present the results of three heterogeneous queries in Figure 5. For each query, the figure shows keyframes of four of the returned shots. The first query describes large-scale motion, such as a group of objects moving from right to left or a camera motion (pan or travelling) to the right.[2] The best match, Figure 5(b), is a tracking shot where the camera passes through under a bridge from left to right. Similarly, the result in Figure 5(c) is a tracking shot where the camera is mounted orthogonally to the direction of travel and captures the passing by environment. The third shot in Figure 5(d) shows a train passing by (from right to left) captured from a static camera. A remarkable result is the fourth shot, Figure 5(e). It captures a large crowd of people which disorderly pushes and shoves to the left.

The second query in Figure 5(f) contains a diagonal motion from left to right which is typical for the film-maker of the analysed films. The best match is a tracking shot, where the camera is mounted approximately 45 degrees to the direction of travel (Figure 5(g)). This yields a dominant motion component in the query vector's direction resulting in a high matching score. The remaining shots show groups of objects (vehicles, people and horses) moving diagonally towards the static camera (Figures 5(h)–5(j)). The motion directions in the returned shots slightly deviate from the query vector's direction. This demonstrates the tolerance of the matching scheme.

The proposed method is able to retrieve even more complex motion compositions. The third

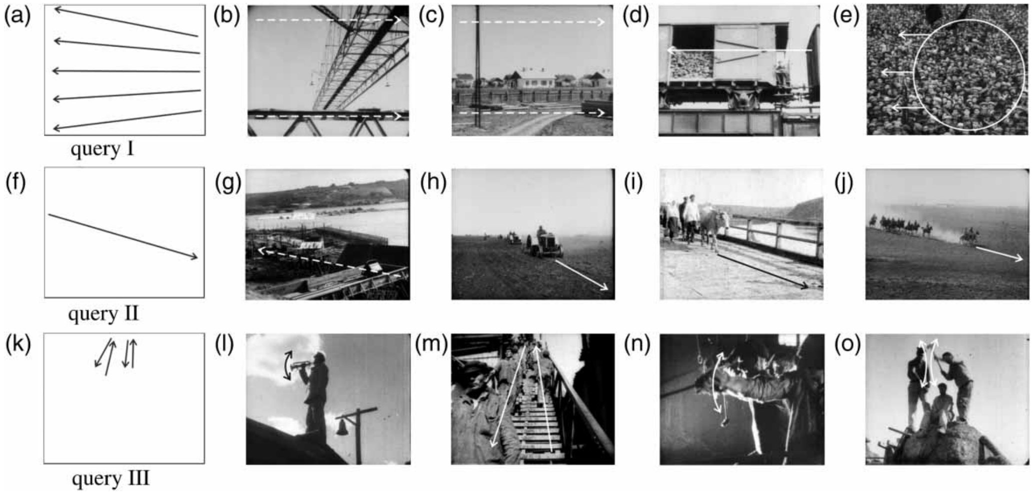

Figure 5. Three example queries. For each query (first column) keyframes of four top-ranked result shots are shown. Dashed arrows in the keyframes mark camera motions and solid arrows are object motions.
Source: Austrian Film Museum; mark-up and illustrations: Matthias Zeppelzauer. Copyright Austrian Film Museum.

query represents a combination of opposed (possibly cyclic) vertical motions in the upper region of the frame. The best match is a shot of a trumpeter who moves his trumpet up and down while playing (Figure 5(l)). The second retrieved shot shows several workers walking up and down a stairway. The shot in Figure 5(n) shows factory workers moving up and down their arms while they pull and push a rod. The last shot in Figure 5(o) shows three workers who hammer down a pin into a rock.

We perform a quantitative evaluation to obtain representative and objective performance measures. For this purpose, we define seventeen different motion queries. The queries represent typical motion patterns of camera motions, small and large objects motions, contrapuntal and rhythmical motion compositions. For each query, we assess the twelve top-ranked returned shots as either relevant or not relevant. The portion of relevant shots in this result set gives a performance measure for the accuracy (precision) of the method and can be computed for differently sized result sets (e.g. from 1 to 12). The corresponding measures are termed “prec@1” to “prec@12”. These measures enable the evaluation of the obtained retrieval performance as well as the ranking generated by the approach. A good ranking is represented by a high prec@1 (which means that the top-ranked result is most often relevant) and monotonically decreasing precisions for larger result sets (prec@2, prec@3, . . .). The average precisions (for result set sizes from 1 to 12) over all seventeen queries are shown in Figure 6. We observe that the precision is generally high and decreases for increasing result set sizes which proves that the ranking is reasonable. The first returned shot is relevant with 94% in average. Among the twelve top-ranked shots of the seventeen queries in average 65% are relevant to the user.

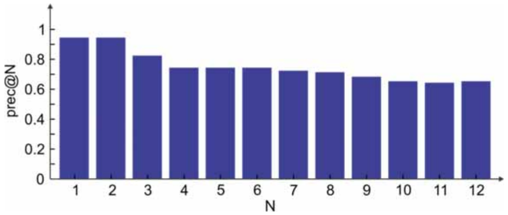

Figure 6. Precisions for different result set sizes.

The evaluation confirms that the generated motion segments adequately represent the motion content of the analysed film material. The simple and intuitive queries combined with the tolerant matching scheme enable the efficient search for particular motion compositions. The method can be further employed in post-production and editing for finding shots with particular motion content, e.g. shots with high/low motion activity, shots with distributed/localised motion, etc. Thus, the method supports the work of both, film creators and film theorists.

3.3 Retrieval of matching actions

Continuity editing plays an important role in filmmaking. It ‘refers to the matching of individual scenic elements from shot to shot so that details and actions, filmed at different times will edit together without error’ (Beaver 2009, p. 59). Continuity editing assures that consecutive shots in a scene fit seamlessly together and that the conveyed story is presented consistently to the viewer. An important device for achieving continuity is *matching on action* (also cutting on action, cutting on motion). Matching on action aims at keeping the screen direction (the motion direction of objects from the perspective of the camera) between successive shots consistent. Directional continuity is important to avoid confusion for the observer. In scenes presenting a chase, for example, the motion direction is usually consistent among several shots, e.g. from left-to-right to convey the impression of a continuous action. A shot that contains right-to-left motion would confuse the observer’s orientation. Two examples of matched actions are shown in Figure 7. The first example in Figure 7(a) and 7(b) shows a character that turns its head from left to right. During this movement the director cuts to a close up of the face. The continued motion between both shots makes the transition between the different shot scales appear seamless. Figures 7(c) and 7(d) show an example of the entrance–exit pattern (Beaver 2009). An actor exits the frame at the right side and in the successive shot enters the

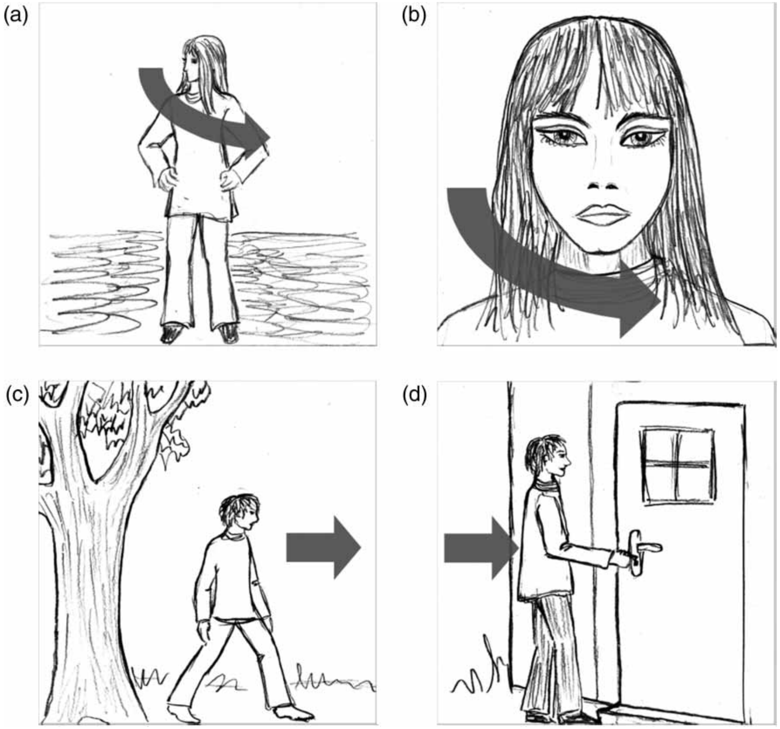

Figure 7. Two examples of matched action.
Source: Krista Zeppelzauer.

frame from the left but perhaps at another time and location. This pattern can be utilised to create a continuous transition between different scenes and locations.

We develop a method for the automatic analysis of continuity editing in a movie. The method finds shots with matching action based on a user-defined query. For this purpose, we extend the presented query model and adapt the matching scheme from Section 3.2. Figure 8 illustrates the retrieval process. First, the query window is split vertically into two halves: A and B. In query A the user sketches the motion present at the end of an arbitrary shot and query B contains the continued motion at the beginning of the next shot. For the retrieval of matched motion we define an analysis window (the grey rectangle in Figure 8) of a few seconds (two seconds in the experiments) around each cut. Query A is then matched *only* with the left half of the analysis window (end of shot N) and query B *only* with the right half (beginning of shot N + 1). The restriction to the window is necessary to get more accurate results. Matching queries A and B with the entire shots N and N + 1 (as in Section 3.2) could take motion into account which is temporally not located around the cut and thus is not relevant for continuity. If both queries positively score on the according halves of the analysis window we combine both scores and return shots N and N + 1 as a result to the user. All returned results are then ranked according to their combined score.

For the retrieval of matched motions a finer matching scheme is necessary that restricts the comparison to the analysis window only. We adapt the matching scheme from Section 3.2 as follows. First, we remove all motion segments that do not coincide with the analysis window.

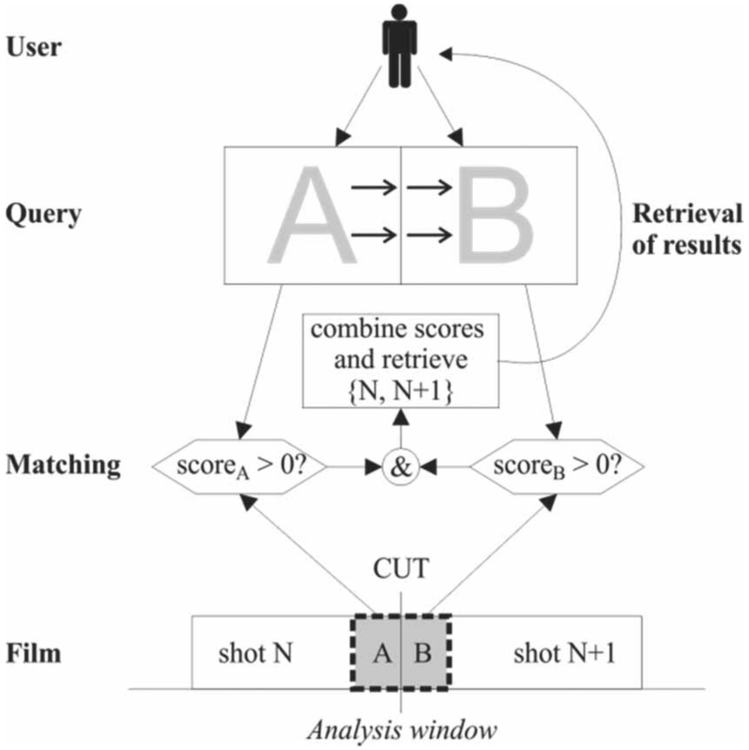

Figure 8. Retrieval of matching actions. Each query (A and B) is matched with the corresponding half of the analysis window. In this example the query represents a typical entrance–exit pattern.

For the remaining motion segments we extract directional and spatial parameters. Since the motion segments may be only partially inside the analysis window, the median direction of the entire segment (as used previously) is not a representative parameter with regard to the analysis window. Instead, we compute the median direction of the segment for *each frame* in the analysis window separately. For an analysis window that covers $2 \cdot D$ frames, this results in D directions $\phi_{j_{1\ldots D}}$ for each segment which allows for a more precise matching. Directional matching between a segment and a query vector is then performed by matching the direction of the query vector with each median direction of the segment. We compute the mean of the cosine similarities (see first term in Equation 2) between the query vector's direction θ_i and each direction ϕ_{j_k} of the segment. Spatial matching is performed as in Section 3.2. The modified scoring function is defined as:

$$s_{i,j} = \frac{1}{D}\left(\sum_{k=1}^{D} \frac{\phi_{j_k} \cdot \theta_i}{\| \phi_{j_k} \| \cdot \| \theta_i \|}\right) \cdot \frac{|R_i \cap M_j|}{|R_i|} \cdot \left(1 - \frac{|M_j \setminus R_i|}{|M_j|}\right). \qquad (2)$$

We compute overall scores for both halves of the analysis window s_A and s_B and combine them by taking their product. This measure yields a high overall score only when *both* scores are high. Additionally, the scores are weighted by the portion of scoring query vectors P_q from queries A and B and by the portion of scoring motion segments P_s in the analysis window: $s_{total} = s_A \cdot s_B \cdot P_q \cdot P_s$.

The weighting increases the score where the query vectors correspond particularly well with the actual motion content. This weighting improves the ranking of the retrieved sequences.

3.3.1 Evaluation

We evaluate the performance of the proposed method with the movie *Run Lola Run* by Tom Tykwer from 1998. The movie is a thriller that makes extensive use of matching on action. There are numerous scenes where motion is consistently carried across several cuts such as chasing scenes and journeys. Many scenes, for example, show the leading character Lola running through the streets from different viewpoints. They are all connected by matching action to create the impression of a continuous journey. Another frequent pattern is subsequent dolly forward shots joined with matching action which show Lola's view during running. The movie further contains characteristic sequences of shots with discontinuous motion. The director connects contrapuntal motions over consecutive shots, for example by alternating dolly forward

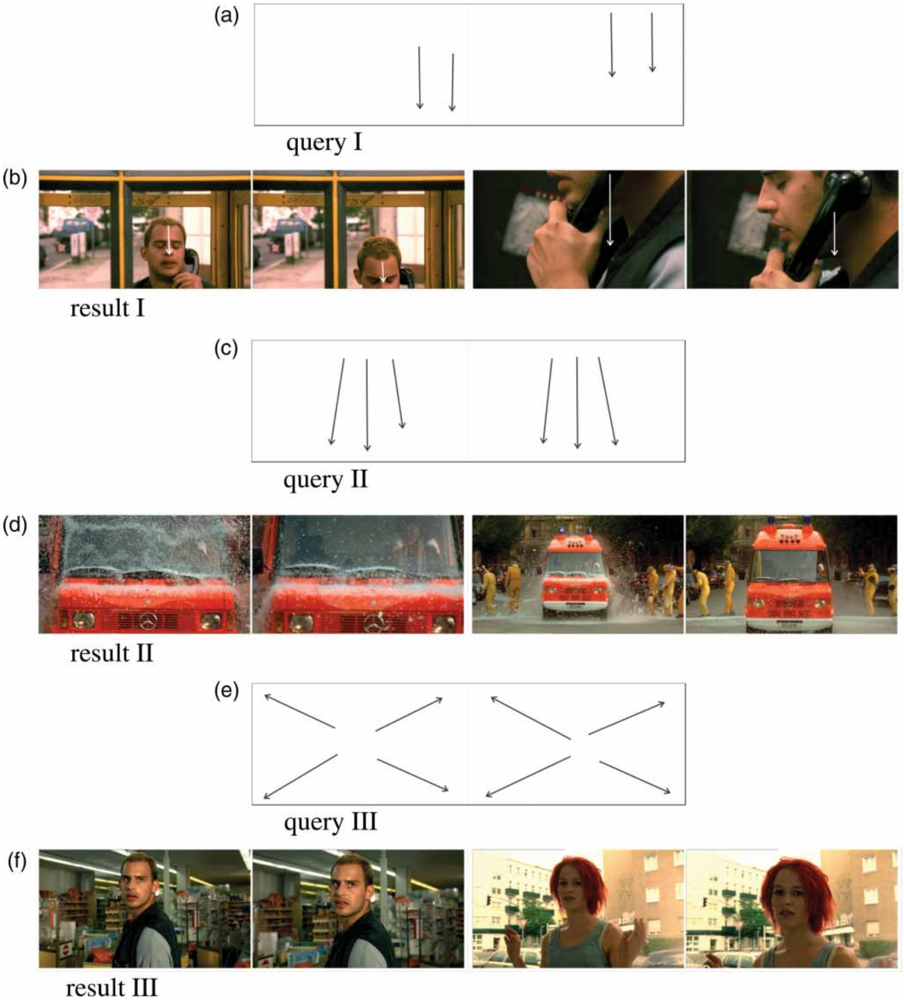

Figure 9. Three example queries that describe different scenarios of matched action together with a corresponding result from *Run Lola Run*.
Credit: Stadtkino Filmverleih; mark-up and illustrations: Matthias Zeppelzauer. Copyright Stadtkino Filmverleih.

and backward movements. The movie is composed of three episodes that show three different versions of the same story. Consequently, for many scenes there exist three different versions with similar motions but varying content. This makes the material well-suited for the evaluation of the proposed approach.

Prior to the evaluation, film experts manually searched for matching actions in the movie and annotated them. We evaluate the retrieval performance for matched actions with different queries. Figure 9 shows results for three example queries. The first one in Figure 9(a) describes a local motion in the right half of the frame directed downwards that is carried across a cut. This is a variation of the entrance–exit pattern. The first returned result shows Lola's friend Manni in a phone booth talking to Lola. At the end of the shot Manni sinks down in resignation and his head leaves the frame at the bottom. The following shot continues this motion from another camera angle and shows Manni's head moving in from the top of the frame.

The second query (see Figure 9(c)) describes a spatially more distributed motion with a screen direction pointing downwards. This can either correspond to downwards motion or motion towards the camera. The query matches well with a scene of shots showing an ambulance approaching the camera that just crashed into a glass plate. While the ambulance comes closer to the camera the director cuts to a wide angle shot that continues the motion (until the ambulance stops) and shows what has happened in the surrounding of the ambulance after the accident (see Figure 9(d)). This example demonstrates how matching action can be applied to create a seamless transition between two different scales of a shot.

The third example query in Figure 9(e) represents the motion pattern produced by a zoom in or dolly forward motion. The method returns several pairs of shots with a continued dolly forward motion. One example is shown in Figure 9(f). The first shot is a medium shot of Manni. While the camera slowly moves towards Manni, the director cuts to a medium shot of Lola where the camera continues its movement at the same speed towards Lola. This transition directs the attention towards the two main characters and increases the tension in the scene.

In most cases the returned results match well with the expert annotations. However, there are also results that do not match the underlying query at the first glance. An example is shots captured with a shaky camera (steadicam) which often appear in chasing scenes. In some cases the shaking of the camera produces a pattern of continued motion between two consecutive shots. Such sequences of shots usually do not convey the impression of a matched action. Other sources of confusion are background movements that match well with the query (e.g. a car moving in the background). Such results may surprise the user because the background motion is often not perceived consciously by the viewer. Consequently, the viewer would not recognise the matching motion in this case. However, the proposed method detects such matching background motions since it does not distinguish between foreground and background motion or between 'salient' and 'non-salient' motion. Although matching background motions may not be recognised by the viewer as examples of matched actions at the first glance, the question arises whether or not the movements are matched on purpose by the director.

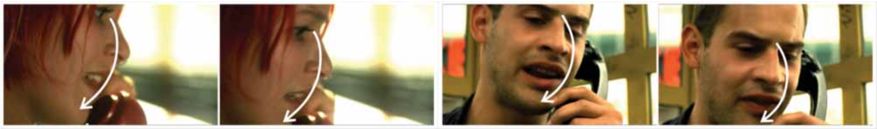

Figure 10. A matched action recovered by the proposed method.
Credit: Stadtkino Filmverleih; mark-up: Matthias Zeppelzauer. Copyright Stadtkino Filmverleih.

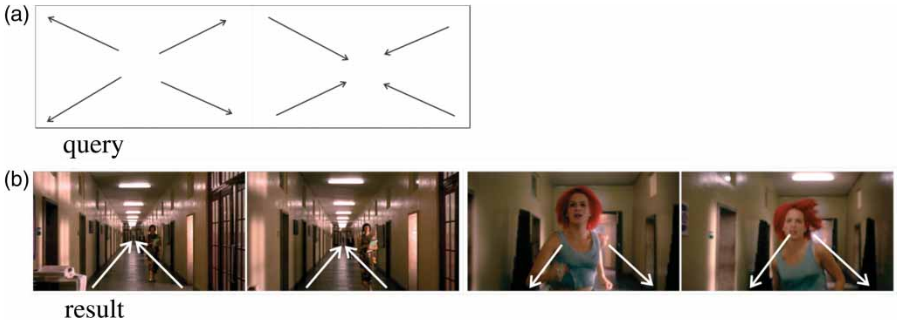

Figure 11. A typical contrapuntal composition of motion that appears repeatedly in the movie together with the according query. *Credit*: Stadtkino Filmverleih; mark-up and illustrations: Matthias Zeppelzauer. Copyright Stadtkino Filmverleih.

We further observe that the proposed method is able to retrieve matching actions that were not recovered by the film experts during annotation. An example is shown in Figure 10. It shows an excerpt of cross cut shots between Manni and Lola talking on the phone. The cut between the shots in Figure 10 is positioned in a way that the downwards movement of Lola's head is continued seamlessly by the downwards movement of Manni's head. This generates a transition that leads the viewer smoothly from one shot to the next. The matching action lasts for approximately a second, which makes it difficult to detect manually. This example shows that the proposed method has the potential to detect patterns of interest that human viewers are likely to overlook. In this manner, the method is able to assist the investigation and exploration of continuity editing in a movie.

The analysis of matched motion with *continuous* screen direction is only one possible application scenario. We can, for example, retrieve sequences of shots with *contrapuntal* motion where shots with different (possibly opposing) motion directions alternate. Such sequences can be employed to create the impression of objects or people moving away from each other. Furthermore, contrapuntal motion is a device for the creation of rhythmic motions. In *Run Lola Run*, for example dolly forward movements are frequently followed by dolly backward movements and vice versa. Retrieval results show that this combination often appears in the movie in situations where Lola is running (see Figure 11 for an example). The first shot of such a combination typically shows the world from Lola's perspective translated in a dolly forward when she is running. In the subsequent shot the camera is positioned in front of Lola and moves backwards showing her running.

Ultimately, we want to point out that the applicability of the method is not limited to the presented examples. The method can be employed for the retrieval of arbitrary combinations of consecutive motions since the definition of the query is up to the user. This makes it well-suited to investigations of film theorists and archivists (e.g. for the study of characteristics of different film genres). Furthermore, the method can be easily extended to automatically retrieve arbitrary matching actions without the need of a query. This functionality could be integrated into automatic production tools such as '*nm2*' which enables the generation of personalised documentaries and summaries based on predefined montage rules (Alifragkis and Penz 2006). The integration of motion analysis would enable the generation of movies with motion continuity and with preferred motion compositions. The stronger incorporation of motion would improve the narration and probably the aesthetic value of the movies generated.

4 Conclusion

This article presents novel methods for the retrieval of motion in movies. Starting with the tracking of feature points and the generation of trajectories, motion segments are created that describe the motion content of a shot in a meaningful way. The connection of robust motion segmentation with an intuitive query model and a tolerant matching scheme enables the retrieval of motion compositions and matching actions. The proposed query model is easy to understand and does not restrict the user to a predefined vocabulary. Consequently, it does not only support searching for already known motion patterns but also enables the exploratory search of motion compositions.

The developed methods perform well in the investigated retrieval scenarios. In some cases, we even gain new insights about the analysed movies by experimenting with the retrieval systems. The presented methods have the potential to assist film analysis and to improve searching movie databases.

In the future we plan to extend the evaluation to a heterogeneous set of film genres, for example action movies which usually contain very fast motions, but also slow-paced genres with less motion activity (e.g. romance films and melodramas). Additionally, we will perform user studies with different user groups (theorists and film-makers) to assess the performance of the proposed methods on a wider basis. A next step on the technical side will be the detection and assembly of arbitrary matching actions (without query) for the generation of movies with motion continuity.

Acknowledgements

We want like to thank Adelheid Heftberger for providing film annotations and film theoretic expertise. This work received financial support from the Vienna Science and Technology Fund (WWTF) under grant number CI06 024.

Notes

[1] Note that for camera motion the direction of the query vectors has to be reversed because the vectors always represent motion relative to the image content. A camera pan to the right, for example, is represented by several arrows that point to the left because the image content on the screen actually moves to the left during a pan to the right.

[2] Note that this query may retrieve camera motions as well as object motions. We do not intend to distinguish between camera and object motion.

References

Ahanger, G., Benson, D., and Little, T.D., 1995. Video query formulation. *Storage and Retrieval for Image and Video Databases III*, 2420 (1), 280–291.

Alifragkis, S. and Penz, F., 2006. Spatial dialectics: montage and spatially organized narrative in stories without human leads. *Digital Creativity*, 17 (4), 221–233.

Ardizzone, E., La Cascia, M., and Molinelli, D., 1996. Motion and color-based video indexing and retrieval, *In: Proceedings of the International Conference on Pattern Recognition*, 25–29 August 1996 Vienna, Austria. IEEE Computer Society, 135–139.

Bashir, F., Khokhar, A., and Schonfeld, D., 2007. Realtime motion trajectory-based indexing and retrieval of video sequences. *IEEE Transactions on Multimedia*, 9 (1), 58–65.

Beaver, F., 2009. *Dictionary of film terms: the aesthetic companion to film art*. 4th ed. New York: Peter Lang.

Bordwell, D. and Thompson, K., 2008. *Film art: an introduction*. 8th ed. New York: McGraw Hill.

Brox, T., *et al.*, 2004. High accuracy optical flow estimation based on a theory for warping. *In*: Tomáš Pajdla and Jiri Matas, eds. *Proceedings on the European Conference on Computer Vision*, 11–14 May 2004 Prague, Czech Republic. Springer, vol. 3024, 25–36.

Buzan, D., Sclaroff, S., and Kollios, G., 2004. Extraction and clustering of motion trajectories in video, *In*: Josef Kittler, Maria Petrou and Mark Nixon, eds. *Proceedings of the International Conference on Pattern Recognition*, 23–26 August 2004 Cambridge, UK. IEEE Computer Society, 521–524.

Chang, S., *et al.*, 1998. A fully automated content-based video search engine supporting spatiotemporal queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 8 (5), 602–615.

Dagtas, S., *et al.*, 2000. Models for motion-based video indexing and retrieval. *IEEE Transactions on Image Processing*, 9 (1), 88–101.

Dimitrova, N. and Golshani, F., 1995. Motion recovery for video content classification. *ACM Transactions on Information Systems*, 13 (4), 408–439.

Hanis, A. and Sziranyi, T., 2003. Measuring the motion similarity in video indexing. *In*: Mirislav Grigic and

Sonja Grigic, eds. *Proceedings of the EURASIP Conference focused on Video/Image Processing and Multimedia Communication*, 2–5 July 2003 Zagreb Croatia. IEEE Computer Society, vol. 2, 507–512.

Horn, B. and Schunck, B., 1981. Determining optical flow. *Artificial Intelligence*, 17, 135–203.

Run Lola Run. 1998. Film. Directed by Tom Twyker. Stadtkino Filmverleih. DVD. X Filme Creative Pool GmbH.

Shi, J. and Tomasi, C., 1994. Good features to track, *In*: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 21–23 June 1994 Seattle, USA. IEEE Computer Society, 593–600.

Zeppelzauer, M., Mitrović, D., and Breiteneder, C., 2008. Analysis of historical artistic documentaries, *In*: *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services*, 7–9 May 2008 Klagenfurt, Austria. IEEE Computer Society, 201–206.

Zeppelzauer, M., *et al.*, 2010. A novel trajectory clustering approach for motion segmentation. *In*: Susanne Boll, *et al.* eds. *Proceedings of IEEE Multimedia Modelling Conference*, 6–8 January 2010 Chongqing, China. Springer, 433–443.

Matthias Zeppelzauer is a PhD student at the Interactive Media Systems Group at the Vienna University of Technology. He received the Master of Science degree in Computer Science in 2005 from the Vienna University of Technology. He has been employed in different research projects focusing on content-based audio and video retrieval. His research interests include content-based retrieval of video and sound, time series analysis, data mining and multimodal media understanding.

Maia Zaharieva is pursuing a PhD at the Interactive Media Systems Group at the Vienna University of Technology, Austria. Her research interests include visual media analysis, processing and retrieval. Zaharieva received her MSc in business informatics from the University of Vienna.

Dalibor Mitrović is a teaching and research associate with the Institute of Software Technology and Interactive Systems at the Vienna University of Technology. He received a Master of Science degree in Computer Science in 2005 from the Vienna University of Technology, Austria. Dalibor Mitrović pursues a PhD in computer science focusing on multimodal information retrieval. His research interests include audio retrieval, real-time feature extraction and computer vision.

Christian Breiteneder is a full professor for Interactive Systems with the Institute of Software Technology and Interactive Systems at the Vienna University of Technology. Christian Breiteneder received the Diploma Engineer degree in computer science from the Johannes Kepler University in Linz in 1978 and a PhD in computer science from the University of Technology in Vienna in 1991. Before joining the institute he was associate professor at the University of Vienna and had postdoctorate positions at the University of Geneva, Switzerland, and GMD (now Frauenhofer) in Birlinghoven, Germany. His current research interests include interactive media systems, media processing systems, content-based multimodal information retrieval, 3-D user interaction, and augmented and mixed reality systems.