Statistical Analysis of Global Motion Chains

Jenny Yuen[1]* and Yasuyuki Matsushita[2]

[1] CSAIL, Massachusetts Institute of Technology, Cambridge MA 02139
[2] Visual Computing Group, Microsoft Research Asia, Beijing 100080, China
jenny@csail.mit.edu, yasumat@microsoft.com

Abstract. Multiple elements such as lighting, colors, dialogue, and camera motion contribute to the style of a movie. Among them, camera motion is commonly overlooked yet a crucial point. For instance, documentaries tend to use long smooth pans whereas action movies usually have short and dynamic movements. This information, also referred to as global motion, could be leveraged by various applications in video clustering, stabilization, and editing. We perform analyses to study the in-class characteristics of these motions as well as their relationship with motions of other movie types. In particular, we model global motion as a multi-scale distribution of transformation matrices from frame to frame. Secondly, we quantify the difference between pairs of videos using the KL-divergence of these distributions. Finally, we demonstrate an application modeling and clustering commercial and amateur videos. Experiments performed show advantage compared to the usage of some local motion-based approaches.

1 Introduction

Global motion plays an important role in defining and characterizing the style of a movie. Think about the style of Alfred Hitchcock's movies, where his characteristic talent manipulating the audience's fears and desires was known greatly for his technique in moving and placing his camera [28, 2]. Another example is the highly dynamic motion in *The Blair Witch Project*, where critics have described how the camera motion can be used to reveal a story through its startling images and lucid characterizations [19]. Moreover, global motion not only varies depending on the director style, but also its content. For example, the characteristic smooth and continuous camera motions in *National Geographic* documentaries are different from the fast and diverse mix of camera movements in an action movie like *Spiderman*.

Videos of different genres clearly contain different *global motion chains*, which is a sequence of global motions. In this work, we study the statistics of global motion *chains* in the spirit of analogous works analyzing image and local flow statistics [23, 11]. Unlike previous works that use global motions for some particular applications such as video classification, video stabilization and etc., this work focuses on the development of a framework to analyze global motion chains in two aspects (Figure 1):

* Part of this work was done while the author was visiting Microsoft Research Asia.

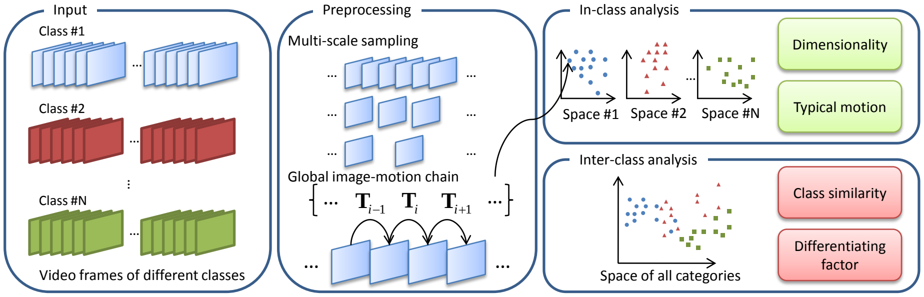


Fig. 1. Overview of the analysis of motion chains.

1. *In-class Analysis*. We want to learn what the motion in a class of videos is like by finding its characteristic camera movements and learning about their intrinsic dimensionality.
2. *Inter-class Analysis*. We want to learn what makes camera motions of a class different from motions in another class and, if particular motions are better at distinguishing a class from another one.

Furthermore, we apply the learnt statistics to compare and cluster videos using KL-divergence as an application of the analyses. We show the results for three types of professional (action movies, basketball games, and documentaries) as well as unprofessional videos taken by numerous people. Finally, we compare the effectiveness of our feature with other motion-based ones when used for clustering and will demonstrate the differences and advantages of our feature in the area of video clustering.

1.1 Related work

There has been a lot of work using global image motion but to the extent to our knowledge, none studying the statistics of global motion chains. The notion of global motion chains, or sequence of transformations between consecutive video frames in a video, has been widely used in the context of video stabilization [13, 20, 16]. These tend to be parameter-based with the objective of smoothing sequences for motion compensation. Motion compensated frames are obtained by warping the original video frames by a cascade of the original and smoothed transformation chains.

While not using the global motion chains, there are several previous works analyzing actions in video clips. These are classified in model-based approaches [21, 18, 1, 14, 24] and nonparametric approaches [5, 6, 30]. The model-based approaches are designed for detecting predefined activities occurring in videos. These are shown to be useful for restricted conditions where the assumed models are valid. More recent trends in action analysis lies in non-parametric approaches.

Chomat and Crowley [5] use a set of spatio-temporal filters and evaluate the joint statistics of filter responses for the purpose of probabilistic action recognition. Efros *et al.* [6] use dense optical flow fields as a feature to perform a nonparametric action recognition. Zelnik-Manor and Irani [30] propose a multi-scale spatio-temporal feature which captures the similarity between video clips based on the behavior contained in the videos. Their feature is based on a local intensity gradient in space-time domain, and it successfully captures local motion occurring in videos. Wang *et al.* [29] developed various methods for detecting and classifying events in surveillance videos using hierarchical linear discriminant analysis (LDA) to cluster local motion attributes. Stauffer *et al.* [27] [9] use tracking for recognition and segmentation of objects in the video . These methods are efficient and robust in many applications such as surveillance or activity recognition and have been extensively tested in still camera settings. We further extend the problem domain and work with videos in a large number of settings and containing significant camera motion.

For the purpose of video classification and other applications, the global motion has been used by several researchers. Roach *et al.* [22] use a background image motion represented by (x, y) translation for the purpose of video genre classification. Affine camera motion is used as one of video features by Smith and Kanade [26] for video skimming. Kobla *et al.* [12] use the camera motion feature for identifying sports videos. Bouthemy and Ganansia [4] propose a method for finding shot changes in a video using global motion. Closest to our work, Fablet *et al.* [7] use global motion for video classification and retrieval. The global motion is represented by a temporal Gibbs random field to perform the temporal analysis. All these works show interesting applications for camera dynamics. We present this paper with a different focus: we want to learn and understand the statistics of motion chains of camera motions across different types of video.

Our work differs from these previous works because we focus on a multi-scale analysis of global motion chains. In particular, we are interested in observing the degree of information contained in the global motion chains across videos of different categories and, as an example, to evaluate to what extent classification can be done in a setting of varied short clips and a moving camera.

While, to the best of our knowledge, there has been no work studying the statistics of global motion chains, Roth and Black [23] propose a model to learn the spatial statistics of optical flow fields using 3D range data. They model the flow field as a *Field-of-Experts* and show numerous statistics of the velocity and orientation of this information. Furthermore, they apply these statistics as a prior for a more robust estimation of optical flow. This paper is analogous to Roth and Black's work, with the objective of learning the form, distribution, and characteristics of global motion chains, which has not yet been done.

1.2 Background

Let $V = \{I_1, ..., I_n\}$ be an image sequence composed of n frames where I_i is its i-th frame. Assume for now that only the camera moves while the environment captured in the video is static and planar. We refer to the term *global motion* between two consecutive frames as a 3×3 projective transformation matrix $\mathbf{T}$. By representing 2D pixel positions (x, y) in homogeneous coordinates $\mathbf{x} = (x, y, 1)^T$, the geometric transformation $I(x, y) \overset{\mathbf{T}}{\Longrightarrow} I'(x', y')$ can be written as $\mathbf{x}' \sim \mathbf{T}\mathbf{x}$, where $\sim$ denotes equality up to scale. The condition for the input video is relaxed for the case of an image sequence with moving objects in the scene by extracting the transformation matrix assuming that the majority of the objects in the captured scene remain static in the real world.

For our global motion estimation, we will assume the affine transformation model. Our reasoning for this assumption is justified by the fact that often the majority of image pixels will be background pixels and they will be far enough from the camera that could be approximated to being on a plane. The affine transformation can be written as:

$$\mathbf{x}' = \mathbf{A}\mathbf{x} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \mathbf{x}, \tag{1}$$

where $\mathbf{x}$ is the original position, $\mathbf{x}'$ is the new position, $\mathbf{A}$ is an affine transformation matrix. This will reduce the number of variables to six per matrix. We also define a vector representation of $\mathbf{T}(= \mathbf{A})$ in scan-line order as $\mathbf{t}$:

$$\mathbf{t} = [a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23}]^T. \tag{2}$$

Global Motion Chains In our analysis, we use ordered *sequences* of global motions which we will refer to as *global motion chains.* We represent the global motion chain of an image sequence centered at the i-th frame as an ordered list of transformation matrices of consecutive frames as:

$$\mathbf{C}_j = \left[[\mathbf{t}_{i-r}]^T, [\mathbf{t}_{i-(r+1)}]^T, \ldots, [\mathbf{t}_{i+r}]^T\right]^T, \tag{3}$$

where r is a positive constant that denotes the radius of the temporal window of the transformation chain $\mathbf{C}_i$, $\mathbf{t}_i$ is the vector representation of the transformation matrix $\mathbf{T}_i$ from frame i to frame $i + 1$.

We further extend this notation to a multi-scale scheme where the sampling of frames is controlled by a frame skip rate s. The global motion chain with the frame skip rate s can be written as

$$\mathbf{C}^s_j = \left[[\mathbf{t}^s_{i-rs}]^T, [\mathbf{t}^s_{i-(r+1)s}]^T, \ldots, [\mathbf{t}^s_{i+rs}]^T\right]^T, \tag{4}$$

where, analogously, $\mathbf{C}^s_i$ is a global motion chain centered at the i-th frame and $\mathbf{t}^s_i$ is the vector representation of the transformation matrix $\mathbf{T}^s_i$ from frame i to frame $i+s$. The frame skip rate allows us to perform a multi-scale representation of the motion information used later in our study.

2 Analysis of Global Motion Chains

In this section we will describe a framework for comparing global motions. We are interested in learning how global motions vary per class and what their

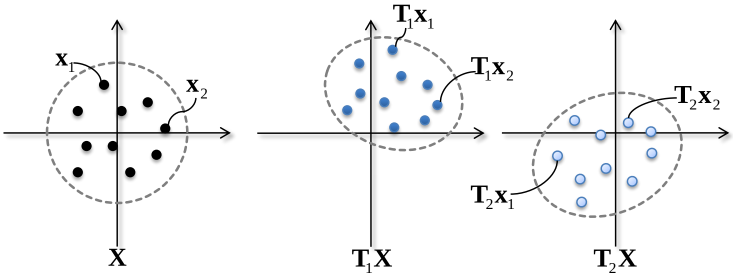

$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{X}$
$\mathbf{T}_1\mathbf{x}_1$
$\mathbf{T}_1\mathbf{x}_2$
$\mathbf{T}_1\mathbf{X}$
$\mathbf{T}_2\mathbf{x}_2$
$\mathbf{T}_2\mathbf{x}_1$
$\mathbf{T}_2\mathbf{X}$

Fig. 2. Illustration of the similarity of global motions. Left: points are distributed in a two dimensional coordinate system. These points are transformed by $\mathbf{T}_1$ and $\mathbf{T}_2$ in the middle and right figures. The similarity of global motions $\mathbf{T}_1$ and $\mathbf{T}_2$ is defined by the sum of distances between corresponding points $\mathbf{X}$.

characterizing elements are. We model a class of videos as a distribution of motion chains. In general, there will be a motion chain over a temporal window r for each frame in each video of a class (excluding some frames at the temporal edges of each video). We will further extend this to a multi-scale approach where we will subsample frames at different rates but the general idea of our model is a *distribution of motion chains.*

2.1 Similarity of global motion chains

We will first define a similarity measure of global motion for general linear transformations in 2-D, e.g., rigid body, similarity, affine and projective. For this case, the similarity of two transformations has the nice and simple property that it can be expressed as the L_2 norm distance between the two transformations when defining the similarity as the amount of point displacement in a 2D plan. The derivation is as follows.

Given some disk in a 2D plane and distributed points on the disk (Figure 2 left). We define the similarity of two linear transformations by the sum of distances of the warped points. Given two linear transformations $\mathbf{T}_1$ and $\mathbf{T}_2$, the similarity $S(\mathbf{T}_1, \mathbf{T}_2)$ of these two transformations can be described as

$$\begin{aligned} S(\mathbf{T}_1, \mathbf{T}_2) &\overset{\text{def}}{=} \frac{1}{|D|}\int_D p(x,y)||\mathbf{T}_1(x \quad y)^T - \mathbf{T}_2(x \quad y)^T||_2^2 dxdy \\ &= \frac{1}{|D|}\int_D p(x,y)||(\mathbf{T}_1 - \mathbf{T}_2)(x \quad y)^T||_2^2 dxdy, \end{aligned} \tag{5}$$

where D represents a set of points distributed on the disk, $|D|$ is the number of points, $||\cdot||_2$ denotes the L_2 norm, and $p(x,y)$ is the probability density function of the point at (x,y). Assuming that $p(x,y)$ has a uniform distribution, as it is the case in a regular 2D images if we consider pixels as points, the above equation can be simplified to

$$S(\mathbf{T}_1, \mathbf{T}_2) = Z||\mathbf{t}_1 - \mathbf{t}_2||_2^2, \tag{6}$$

where the constant Z is written as

$$Z = \frac{1}{|D|} \int_D p(x,y) ||(x \quad y)^T||_2^2 dxdy. \tag{7}$$

From this result, the similarity of two linear transformations can be computed by the L^2 norm. Consequently, the similarity $S_{\mathcal{C}}$ of a pair of global motion chains $\mathcal{C}$ and $\mathcal{C}'$ can be written as

$$S_{\mathcal{C}}(\mathcal{C}, \mathcal{C}') = \sum_i S(\mathbf{T}_i^{(\mathcal{C})}, \mathbf{T}_i^{(\mathcal{C}')}), \tag{8}$$

where $\mathbf{T}_i^{(\mathcal{C})}$ represents the i-th transformation matrix in $\mathcal{C}$.

The similarity metric can be augmented by modifying the probability density function $p(x,y)$. For example, if we assume non-uniform importance of pixels, i.e., regions of interest, $p(x,y)$ becomes non-uniform. While it is possible to use this generalized form of the similarity metric, we assume the uniform density of $p(x,y)$ for simplicity.

2.2 In-class Analysis

In-class typical motions The purpose of this section is to study what typical motion chains are like. Intuitively, we can think about how shaky camera motion could characterize handheld-captured videos, slow pans might be common in documentaries about nature, or highly dynamic movements in action movies.

To identify the typical transformation chains, we fit a mixture of Gaussians for each class distribution of transformation chains. The means of the Gaussian centers describe *typical* transformation chains in the class. We choose the number of Gaussians m according to the Bayesian information criterion (BIC) [25]:

$$m = \underset{m}{\operatorname{argmin}} \{-\log P(data|m, \hat{\theta}_m) + \frac{k_m}{2} \log n\}, \tag{9}$$

where k_m is the number of parameters given m Gaussians, $\hat{\theta}_m$ is the maximum likelihood parameters for the distribution with m Gaussians, and n is the number of points in the distribution.

Space of global motion chains We will also study the compactness in the distribution of each class of videos. For example, consider the extreme case for surveillance videos where the camera is static. In this case, the camera motion of this video can be compactly characterized by this static state (or viewed in our terms as an identity matrix since there is no global motion from frame to frame). On the other hand, if we have completely random motion, the information required to characterize the space will be larger than in the previous example.

To analyze the compactness of the in-class distribution of transformation chains, we project each distribution of points to their principal components and analyze the cumulative variance of the distribution as a function of the number of principal components selected.

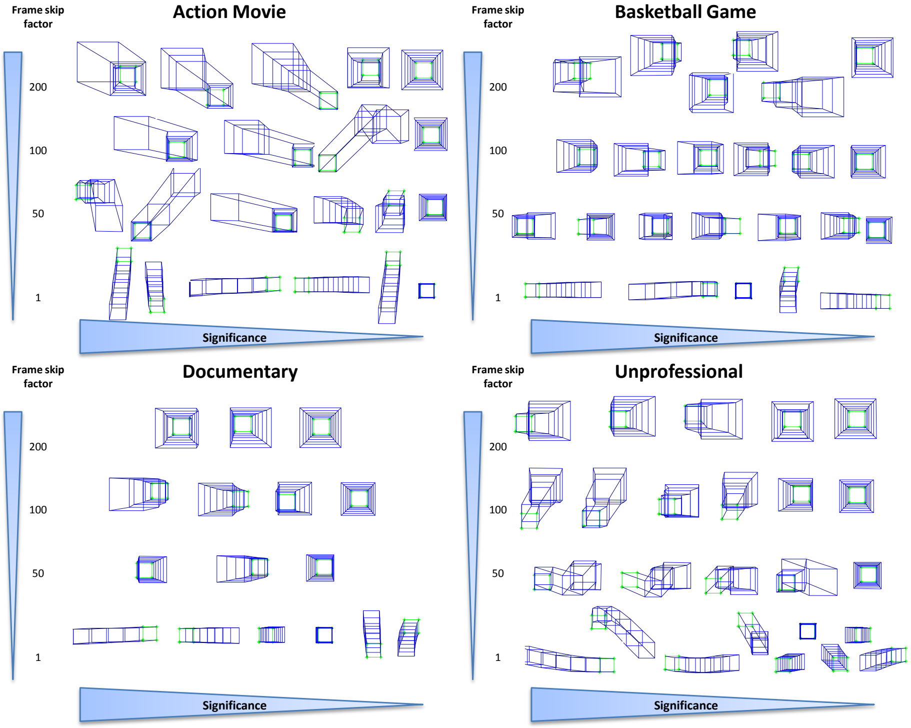


Fig. 3. Characteristic motions in each category of videos across different frame skip rates. The motions are ordered by descending significance (see Section 2.3) from left to right. Each frame in the motion chain is described as a rectangle undergoing an affine transformation from frame to frame. The first frame is denoted by the rectangle with the * in each corner.

2.3 Inter-class Analysis

This section focuses on the development of a comparison metric between camera motion distributions. Based on the modeling of a group of movies of a particular class as a distribution of motion chains, we develop a metric to compare between two videos (or groups of videos). Moreover, exploiting the form of a Gaussian Mixture and the representation of a typical motion as a Gaussian component, we propose a metric for ranking characteristic motions in order of *significance*, which intuitively means we order typical motions with respect to how useful they are distinguishing the class they belong to from others.

Video Comparison The divergence of two distributions is useful for comparing pairs of video classes, which in turn can be used for clustering of individual videos. In section 4 we show how we apply this metric for video clustering.

We compare pairs of distributions (each representing a class) and analyze how much one diverges from the other using an approximation of the KL-divergence of two mixtures of Gaussians proposed by Goldgerber *et al.* [8]:

$$KL(f||g) \approx \sum_{i=1}^{n} \alpha_i \min_j \left(KL(f_i||g_j) + \log \frac{\alpha_i}{\beta_j} \right) , \tag{10}$$

where α_i and β_j are the mixing coefficients for f_i and g_j respectively and $KL(f_i||g_j)$ is the KL divergence approximation for two Gaussian mixture components f_i and g_j:

$$KL(f_i||g_j) \approx \frac{1}{2}\left(\log \frac{|\Sigma_{g_j}|}{|\Sigma_{f_i}|} + \mathrm{Tr}(\Sigma_{g_j}^{-1}\Sigma_{f_i}) + (\mu_{f_i} - \mu_{g_j})^T \Sigma_{f_i}^{-1}(\mu_{f_i} - \mu_{g_j}) \right). \tag{11}$$

Significance of a Typical Motion Identifying the canonical motions that differentiate a class from others can be helpful in creating stronger classifiers. In this section we present a ranking method for typical motions to identify what are the motions that characterize each class and at the same time differentiate it from the rest.

As a consequence of equation (10), Goldberger *et al.* [8] propose a method for matching some Gaussian component from a mixture of Gaussians (MoG) f to another component $\pi(i)$ of another MoG g via the following matching function:

$$\begin{aligned} \pi(i,g) = \operatorname*{argmin}_j \Bigg(& \frac{1}{2}\big(\log \frac{|\Sigma_{2,j}|}{|\Sigma_{1,i}|} + \mathrm{Tr}(\Sigma_{2,j}^{-1}\Sigma_{1,i}) + \\ & (\mu_{1,i} - \mu_{2,j})^T \Sigma_{2,j}^{-1}(\mu_{1,i} - \mu_{2,j})\big) - \log \beta_j \Bigg). \end{aligned} \tag{12}$$

Based on this matching-based approximation, we propose a way of ranking canonical motions of each distribution in a set of MoG G as

$$\begin{aligned} p(i) = \min_{g \in \{G-f\}} \Bigg(& \frac{1}{2}\big(\log \frac{|\Sigma_{2,j_g}|}{|\Sigma_{1,i}|} + \mathrm{Tr}(\Sigma_{2,j_g}^{-1}\Sigma_{1,i}) + \\ & (\mu_{1,i} - \mu_{2,j})^T \Sigma_{2,j_g}^{-1}(\mu_{1,i} - \mu_{2,j_g})\big) - \log \beta_{j_g} \Bigg). \end{aligned} \tag{13}$$

This ranking is based on the matching-based approximation from equation 13 where, the larger a value is, the more divergent the typical motion (Gaussian component) is from other typical motions in the set of distributions (other video classes).

3 Experiments and Results

3.1 Dataset and Data Processing

We perform our analyses over two video datasets:

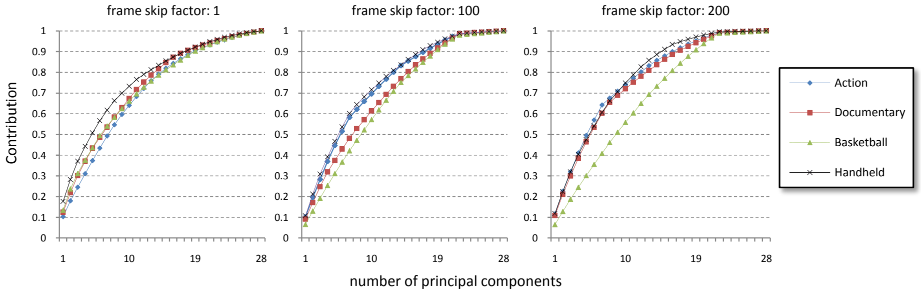


Fig. 4. Plots of cumulative variance as a function of principal components of motion chains. These plots show the size of space accounted by some number of principal components.

1. Dataset 1. Consists of commercially produced video and unprofessional video (home videos captured by various amateurs) divided into the following categories: action movie (180 minutes) , documentary (120 minutes), basketball game (120 minutes), and unprofessional (360 minutes).
2. Dataset 2. Consists of 176 video clips downloaded from YouTube: 130 professional and 46 unprofessional

We extracted the similarity transformation matrices between each pair for consecutive video frames in the sequences sampled at the particular skip rates to construct transformation chains of a temporal window of radius $r = 7$ using the hierarchical Lucas-Kanade algorithm [15, 3]. Using the similarity transformation we have 6 variables, therefore each transformation chain is a feature point in 42 $(= 6 \times 7)$ dimensions in our case.

3.2 In-class Analysis

Figure 4 shows the cumulative variances of the unprofessional videos and the three professional classes of videos as a function of number of principal components used (ordered by eigenvalue). Each graph shows this relation across different levels (skip rates).

In these plots, it is observed that the cumulative variance of the unprofessional dataset (handheld) is consistently high among the four datasets. This suggests that the space of unprofessional motion-chain is more compact compared with others in these levels. This result is counter-intuitive because the unwanted high-frequency motions in unprofessionally-created movies are expected to create a larger variability. One interpretation for this result is that professionally-created movies have wider variety of global motions that are well-designed, while the unprofessional ones do not. This differentiates the variety of global motions, and it appears as a difference in size of spaces.

Figure 3 shows the motion means from each of the unprofessional videos as well as the three professional subclasses. Each row shows characteristic motions at a particular skip rate and in each row, the motions are ordered in order of significance.

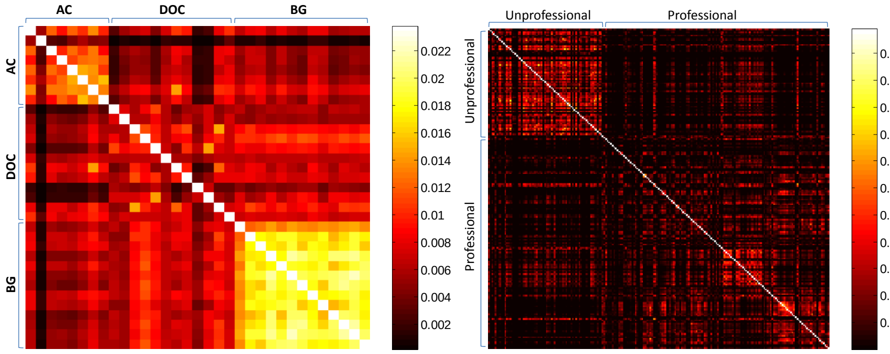
3.3 Inter-class Analysis

Figure 3 shows typical motions in each class ordered by significance (see equation 14). As expected, in all classes, the static camera motion chain was ranked the lowest in terms of significance. In other words, a non-moving camera shot exists in videos of all types and contributes little in characterizing the video or its class. In contrast, more complex motions serve to characterize and discriminate within classes:

- **Action Movies.** Characterized by a variety of horizontal and vertical motions. While some translations are carefully-generated continuous paths, there are some drastic changes in location and zooms. At a low level there are simple vertical and horizontal motions while at a higher skip rate, there are more zooms with highly dynamic motions. For instance, the camera can be focusing on a character and moving quickly when the focus changes to another character.
- **Basketball Games.** Contain very carefully guided curved camera paths with continuous zoom. Low level motions are general motions in horizontal and vertical directions as well as the lack of movement (static camera). High-level motions are mostly careful zooms with little translations. There are more horizontal motions at higher skip rates compared to other classes. This is probably due to the rapid and varied motions of players in a sport of this nature. Moreover, these effects are possible thanks to equipment not available in other classes like the unprofessional one.
- **Documentaries.** Motions are very simple and there are fewer characteristic translations with respect to the other classes studied. At a low level there are continuous motions vertically and horizontally while more complex motions happen at higher skip rates including zooms and carefully controlled paths. At very high skip rates, there are mostly zooms with non or very little translation. This is characteristic of documentaries when filming packs of animals in aerial views.
- **Unprofessional.** This class has several characteristic motions in all directions. Due to the limited translation capability of the cameraman in at a higher level, there is little translation at higher skip rates. In general there are less continuous zooms and unstable translations compared to the professionally-generated videos.

4 Application to video classification

We are interested in seeing to what extent clustering can be achieved only using global motion as a feature. We represent each video clip as a distribution of global motion chains and use the symmetric version of the KL-divergence approximation previously described in section 2 to compare pairs of global motion



(a) Within professional videos (b) Professional and unprofessional videos

Fig. 5. Similarity matrices: (a) Similarity matrix for professional videos: action movies (AC), documentaries (DOC), and basketball games (BG), and (b) Similarity matrix for professional and unprofessional YouTube videos.

chain distributions to create an affinity matrix used to perform a normalized-cuts spectral clustering [17] . The similarity between two video clips with distributions of motion chains f_s and g_s at scale s respectively is therefore defined as follows:

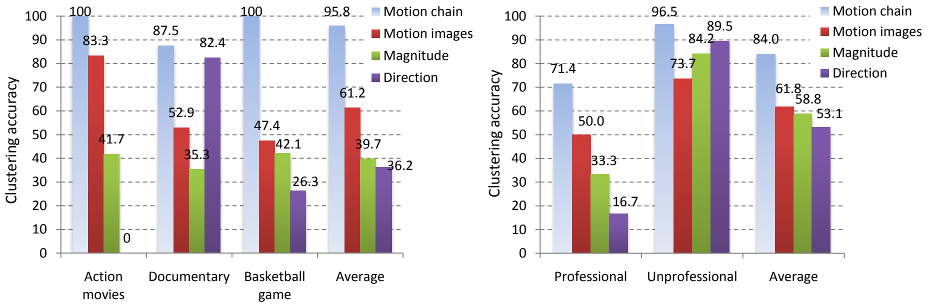
$$S(v_f, v_g) \overset{\text{def}}{=} \frac{1}{\sum_{s=1}^{M} \{KL(f_s||g_s) + KL(g_s||f_s)\}}. \tag{14}$$

Figures 5 (a) and (b) show the affinity matrix for the professional videos in dataset 1 and for all of dataset 2 respectively. The brightness of the region is directly proportional to the affinity of the videos in that section. In dataset 1, basketball game clips are more similar to each (and easier to cluster) other than action movie and documentary clips. In the case of dataset 2, unprofessional clips are more similar to each other than professional videos.

Comparison to other Motion Features We evaluate the effectiveness of our global motion chain feature representation in the dataset described in section 3. The 13 hours of video are cut into 3 minute long non-overlapping clips totalling 253 video clips.

We compare our motion chains feature with three other motion-based features: motion images described by [10], motion magnitude histograms, and motion direction histograms. Motion Images are 2D grids accumulating motion information at the corresponding parts of the canvas where the motion happened. The motion magnitude and direction histograms summarize the magnitude and direction respectively of optical flow vectors throughout the video clip.

We implement a version of motion images and the magnitude and direction histograms using optical flow over pairs of consecutive frames using the block matching implementation in OpenCV. The flow vector direction and magnitude



(a) Dataset 1

(b) Dataset 2

Fig. 6. Clustering accuracy over datasets 1 and 2.

information are summarized in normalized histograms and the values of each bin are used as feature values.

To compute the accuracy of the clustering results within each class, we calculate the ratio between the number of instances that were grouped together (if there is more than one grouping, we choose the largest grouping) and the total number of instances in the ground-truth of that class. Each of the 4 types of features were used to construct similarity matrices and spectral clustering was applied to them. Figure 6 (a) shows the classification rates amongst professional videos using the different features for dataset 1. Global motion performed far better than our competition in all the three classes with an average of advantage of at least 34.57%. The most challenging class was the documentaries one (DOC) where we achieved an 87% accuracy, still significant in comparison to the other methods tested.

Figure 6 (b) shows the performance of our feature when used to discriminate between professional and unprofessional videos in our dataset of YouTube videos. We are able to correctly group unprofessional videos with 96.48% accuracy and professional ones with 71.43% accuracy.

5 Discussion and Conclusion

We perform an analysis of the global motion distribution within various video classes and show how the space of unprofessional motions is more compact with respect to professionally-captured ones at a high level. At a low level, action movies are the least compact followed by basketball games and documentary films in that order. In addition, we learned that unprofessional movies have similar wide range of motion chains with respect to professional films while at a higher level, the camera motion of professional films is more controlled and less varied in comparison with unprofessional movies.

We also develop a way to compare camera motions which, as a result, we apply in a video clustering application. Our method is more efficient at clustering videos when compared to other motion-based features in the literature. Global motion features perform better in our dataset where there is a large variety of camera motions characterizing the genre of the video while other features such as global motion and other low-level ones based on local motion perform well in certain scenarios where the clip is short, the camera is static, and contains simple actions.

In conclusion, have performed a camera motion study and show that camera motion is indeed different across video classes and describe characteristics across classes and what differentiates them. We also show experimentally that global motion carries significant information about the movie type by comparing our feature to other features based on local motion, where the clustering results were consistently better using global motion as a feature. Other potential applications of this work include camera motion transfer between sequences to change styles, recognition of directors in videos, or simple characteristic motion extractors for learning about different genres or direction styles.

References

1. M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.
2. S. Blanford. *The Film Studies Dictionary.* Oxford University Press, 2001.
3. J. Y. Bouguet. Pyramidal implementation of the lucas kanade feature tracker : description of the algorithm. In *OpenCV Document, Intel, Microprocessor Research Labs*, 2000.
4. P. Bouthemy and F. Ganansia. Video partitioning and camera motion characterization forcontent-based video indexing. In *Proceedings of International Conference on Image Processing*, volume 1, pages 905–908, 1996.
5. O. Chomat and J. L. Crowley. Probabilistic recognition of activity using local appearance. In *Proceedings of Computer Vision and Pattern Recognition*, pages 104–109, 1999.
6. A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of International Conference on Computer Vision*, volume 2, pages 726–733, 2003.
7. R. Fablet, P. Bouthemy, and P. Pèrez. Statistical motion-based video indexing and retrieval. In *Proceedings of RIAO Conference on Content-Based Multimedia Information Access*, 2000.
8. J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Proceedings of International Conference on Computer Vision*, 2007.
9. E. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *Proceedings of Computer Vision and Pattern Recognition*, volume 2, pages 22–29, 1998.
10. A. Haubold and M. Naphade. Classification of video events using 4-dimensional time-compressed motion features. In *Computer Vision and Image Understanding*, pages 178–185, 2007.
11. J. Huang. Statistics of natural images and models. In *Ph.D. Thesis*, 2000.

12. V. Kobla, D. DeMenthon, and D. Doermann. Identifying sports videos using replay, text, and camera motion features. In *Proceedings of the SPIE conference on Storage and Retrieval for Media Databases*, pages 332–343, 2000.
13. A. Litvin, J. Konrad, and W. C. Karl. Probabilistic video stabilization using kalman fitering and mosaicking. In *Proceedings of IS&T/SPIE Symposium on Electronic Imaging, Image and Video Communications*, pages 663–674, 2003.
14. F. Liu and R. W. Picard. Finding periodicity in space and time. In *Proceedings of International Conference on Computer Vision*, pages 376–383, 1998.
15. B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
16. Y. Matsushita, E. Ofek, X. Tang, and H.-Y. Shum. Full-frame video stabilization. In *Proceedings of Computer Vision and Pattern Recognition*, volume 1, pages 50–57, 2005.
17. M. Meila and J. Shi. A random walks view of spectral segmentation. In *Artificial Intelligence and Statistics*, 2001.
18. S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in xyt. In *CVPR*, pages 469–474, 1994.
19. T. A. Ohanian and M. E. Phillips. *Digital Filmmaking: The Changing Art and Craft of Making Motion Pictures.* Focal Press, 2000.
20. M. Pilu. Video stabilization as a variational problem and numerical solution with the viterbi method. In *Proceedings of Computer Vision and Pattern Recognition*, volume 1, pages 625–630, 2004.
21. R. Polana and R. C. Nelson. Detecting activities. In *Proceedings of Computer Vision and Pattern Recognition*, pages 2–7, 1993.
22. M. J. Roach, J. D. Mason, and M. Pawlewski. Video genre classification using dynamics. In *ICASSP*, pages 1557–1560, 2001.
23. S. Roth and M. J. Black. On the spatial statistics of optical flow. *International Journal of Computer Vision*, 74(1):33–50, 2007.
24. P. Saisan, G. Doretto, S. Soatto, and Y. N. Wu. Dynamic texture recognition. In *Proceedings of Computer Vision and Pattern Recognition*, volume 2, pages 58–63, 2001.
25. G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 5(2):461–464, 1978.
26. M. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding techniques. In *Proceedings of Computer Vision and Pattern Recognition*, pages 775–781, 1997.
27. C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.
28. M. Walker. *Hitchcocks's Motifs.* Amsterdam University Press, 2005.
29. X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical bayesian models. In *Proceedings of Computer Vision and Pattern Recognition*, volume 2, pages 1–8, 2007.
30. L. Zelnik-Manor and M. Irani. Statistical analysis of dynamic actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1530–1535, 2006.