MEAN SHIFT BASED NONPARAMETRIC MOTION CHARACTERIZATION

Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{lingyu, xumin, tian, xucs}@i2r.a-star.edu.sg

ABSTRACT

Motion content is a very powerful cue for organizing video data. Efficient and robust identification of the camera motion nature and the dominant object motion is important for generation of useful motion annotations. Most of existing methods focus on the estimation of a parametric motion model from dense optical flow fields or block-based MPEG motion vector fields (MVF). However, it is hard to achieve reliable model estimation in large amounts of video data. This is due to the violation of parametric assumption in the presence of large object motion and bad estimation of the optical flow in low-textured regions. In this paper, we employ the mean shift procedure and the histogram to propose a novel nonparametric motion representation. With this motion representation, we transform the motion analysis to the classification problem of camera motion patterns in the presence of dominant object motion and non-dominant object motion. The unique features include three main aspects: 1) Instead of computationally expensive and vulnerable parametric regression we base the motion characterization on the classification of motion patterns, 2) we employ machine learning to capture the knowledge of recognizing camera motion patterns from bad motion fields, and 3) with the mean shift filtering the proposed motion representation elegantly considers the spatial-range cues so as to remove noise and implement discontinuity preserving smoothing of motion fields. Promising results are achieved on 1096 motion vector fields extracted from compressed broadcast soccer video.

1. INTRODUCTION

Motion understanding is a very rich and complicated problem in computer vision. Most of the literatures have examined the problem of parametric or dense motion field estimation for video compression and coding. In the context of content-based video indexing, the primary task generally consists in segmenting the video into elementary shots. Since motion is an integral part of a shot, an interpretation and representation of shot content is usually associated with the recognition of typical forms of video shooting such as static shot, tracking, zooming, and panning. Further dynamic shot content analysis includes image mosaicing, segmentation, tracking and characterization of moving objects.

Dominant motion estimation is the main task of motion analysis for motion-based video indexing and retrieval. Existing works [1-3] usually assume that this dominant motion corresponds to the apparent motion induced by camera motion. A 2-D affine or quadratic model is introduced to model the transformation between two successive images. To estimate the dominant motion, robust statistics is employed to minimize the displaced frame difference. Secondary motions (e.g. object motion) are then considered as outliers in terms of statistics.

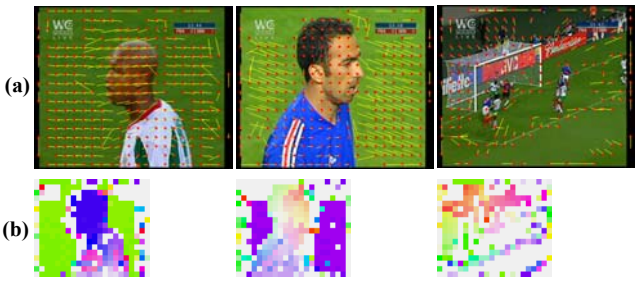

Fig.1. Motion vector fields extracted from MPEG-1 compressed soccer video: (a) original frames overlapped with motion vectors, (b) the motion vector field representations in MVS space [4].

The dominant assumption is closely related to camera shots. The long shot is the most basic of all movie shots. It includes everything of importance in a given scene. The dominant motion is naturally induced by camera movement for a long shot. The close-up concentrates exclusively on one person's face, or on any one detail of a scene. The apparent motion is attributed to the object motion and the camera movement. The assumption of dominant motion induced by the camera movement is usually weak for a close-up. This frequently occurs in a mid shot also. It is easy to spot from the close-up MVF in Fig.1. We thus cannot expect reliable results directly by applying the parametric model estimation to the real motion fields.

Apart from the dominant assumption, dominant motion analysis depends on the optical flow estimation, which is usually based on gradient methods or block matching methods. The quality of optical flow fields affects the performance of motion analysis to a large extent. This is prominent when we directly exploit the MVF extracted from MPEG compressed video. The long shot MVF in Fig.1 corresponds to zooming out. However, the identification of zooming out is a demanding job due to random or non-motion vectors in large low-textured field regions as shown in Fig.1.

Although motion characterization has been examined in lots of literature [3] [5-8] in the context of video indexing, existing works did not explicitly consider the cases when the camera motion do not dominate and/or the complete MVF cannot be recovered by simple filtering (e.g. median filtering). Moreover, most work did not intentionally perform experiments on various motion fields from different camera shots (long shot, mid-shot, and close-up).

0-7803-8554-3/04/$20.00 ©2004 IEEE.

In this paper, we will present a nonparametric method for motion characterization. Currently we are concerned about four tasks: 1) the identification of prominent object (e.g. a player, a referee in sports video) motion (POM), 2) the identification of prominent camera motion (PCM) in the presence of POM, 3) the recognition of typical forms of video shooting in the absence of POM, and 4) the detection of random MVF (RMF) due to an inappropriate optical flow estimation in large low textured regions. The proposed method will treat motion characterization as the classification problem of motion patterns. It does not rely on any parametric motion model or any dominant assumption. As a motion field contains large amounts of motion vectors, it is infeasible to directly feed a MVF into a learning algorithm. We thus employ the mean shift procedure [9] and the histogram to develop a novel nonparametric motion field presentation. With this representation, lots of motion analysis work can be based on the full-fledged machine learning algorithms. Its effectiveness will be demonstrated from the performance about the four tasks.

This paper is organized as follows. In Section 2, we present an overview of this nonparametric motion characterization scheme. In Section 3, we discuss the motion field representation. Around those four tasks we briefly introduce the motion characterization target in Section 4. Experimental results are given in Section 5. Finally, we conclude this paper in Section 6.

2. OVERVIEW

Fig.2 illustrates the proposed motion characterization scheme.

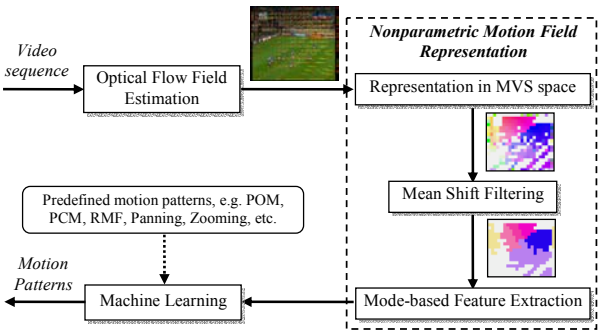

Fig.2. Nonparametric motion characterization scheme

As shown in Fig.2, this scheme includes three main stages: optical flow estimation, nonparametric motion representation, and machine learning. Our main contributions are the mean shift based nonparametric motion representation and the idea of using machine learning mechanism to capture meaningful knowledge of recognizing various motion patterns.

The above-mentioned knowledge in general comprises two parts: domain-independent and domain-dependent. The first includes the dominant assumption, 2-D parametric motion model and so on; the latter includes the influence of different camera shots on motion analysis, potential regions with good motion estimation (e.g. an audience region usually exhibits more robust motion estimation than a low-textured field region in broadcast soccer video) and so on. The latter is practically significant for improving performance. However, it is very difficult to employ some parametric models to represent the domain-dependent knowledge. Therefore, we resort to machine learning algorithms.

In order to avoid the vulnerable dominant assumption and the computationally expensive parametric estimation, we introduce a nonparametric method to represent motion fields. Based on this representation, machine learning is exploited to acquire both domain-independent and domain-dependent knowledge.

We employ the mean shift procedure and the histogram to make the nonparametric motion representation. In general a motion field is represented as a two-dimensional lattice of p-dimensional vectors. The space of the lattice is known as the *spatial* domain, while the motion (magnitude and direction) is represented as the *range* domain. Thus an effective motion representation has to take account of the spatial-range constraints. As the mean shift procedure exhibits excellent discontinuity preserving filtering and elegant mode-seeking functions, we employ it to smooth noisy motion fields and seek representative motion modes. This is different from traditional pre-preprocessing [7] such as noise-removal filtering (e.g. median filtering). Moreover, we use the histogram to represent the spatial distribution for each mode. Cross-validation is exploited to determine appropriate histogram bin widths.

3. NONPARAMETRIC MOTION FIELD REPRESENTATION

As shown in Fig.2, this motion field representation comprises three stages: representation in MVS space, mean shift filtering, and mode-based feature extraction. A brief review of the mean shift procedure is first given. We then describe those stages.

3.1 Mean Shift Procedure

Given n data points $X_i, i = 1,\ldots,n$ in the d-dimensional space R^d, the multivariate kernel density estimator with kernel $K(x)$ and window width h is given by

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left\{\frac{1}{n}(x - X_i)\right\} \tag{1}$$

For the Epanechnikov kernel, the density gradient estimate becomes

$$\hat{\nabla} f(x) = \frac{n_x}{n(h^d c_d)}\frac{d+2}{h^2}\left(\frac{1}{n_x}\sum_{X_i \in S_h(x)}[X_i - x]\right) \tag{2}$$

where the region $S_h(x)$ is a hypersphere of radius h having the volume $h^d c_d$, centered on x, and containing n_x data points. The last term

$$M_h(x) \equiv \frac{1}{n_x}\sum_{X_i \in S_h(x)} X_i - x \tag{3}$$

is called the sample mean at $x \in X$. The repeated movement of data points to the sample means is called the *mean shift procedure* [9]. The mean shift vector always points towards the direction of the maximum increase in the density. Since efficient mean shift computation requires efficient range searching, Comaniciu et al. [10] proposed a computational module of the mean shift procedure, and successfully applied it to two low-level vision tasks: discontinuity preserving filtering and image segmentation.

3.2 Representation in MVS space

According to motion vector characteristics and HSV parameter ranges, we propose a cone-shaped MVS space in [4]. The MVS space provides a visualized representation of the MVF. *Hue* represents motion vector direction and *saturation* motion vector magnitude. Such representation is employed as a visual aid to understand and analyze the characteristics of motion fields. Several illustrative examples are given in Fig.3. More details of converting a motion vector into MVS space can be found in [4].

3.3 Mean Shift Filtering

We employ the mean shift procedure to seek motion modes in the spatial-range joint domain. Once the representative modes are found, the motion field representation in MVS space is transformed to a "mosaic" comprising a set of colored pieces (See Fig. 3). The recognition problem of a motion pattern is to learn the spatial-range composition knowledge of colored pieces for each predefined motion pattern from training samples. To some extent we can think of the mean shift filtering as a kind of dimension reduction under the spatial-range constraints.

We employ the joint domain kernel

$$K_{h_s,h_r} = \frac{C}{h_s^2 h_r^3} k\left(\left\|\frac{X^s}{h_s}\right\|^2\right) k\left(\left\|\frac{X^r}{h_r}\right\|^2\right) \tag{4}$$

to perform mean shift clustering on the representation in MVS space, where X^s and X^r are the spatial part and range part respectively, h_s and h_r are the kernel bandwidths, C is the normalization constant. According to the clustering results, a motion vector field MVF can be represented by

$$MVF = \{P_i\}_{i=1,\ldots,m},\ P_i = < S_i, \overline{MV}_i, R_i > \tag{5}$$

, where S_i denotes the set of motion vectors associated with the cluster P_i, $\overline{MV}_i$ is the average motion vector of S_i, R_i is the normalized cluster size of P_i, $\sum_{i=1}^{m} R_i = 1,\ 0 \le R_i \le 1$.

In Equation (5), each cluster P_i corresponds to a motion mode and each set of motion vectors S_i forms a colored piece. We then perform feature extraction according to the result in (5).

3.4 Mode-based Feature Extraction

To construct a uniform feature vector for all MVFs, we have to solve two problems: 1) an elimination of the variability due to variable modes number in different MVFs, and 2) an effective representation of mode-related spatial information.

We want to select a set of modes from the results in Equation (5) for feature extraction. The criterion of mode selection is based on the priority determined by the cluster size R_i. The mode with a larger size R_i is the first to be chosen. The number of selected modes N_s is predefined. If the cluster number m in Equation (5) is less than N_s, we will employ padding with zero to extend the feature vector to a required dimension.

We employ the histogram to represent spatial distribution for a chosen mode. Let Γ be a $n_1 \times n_2$ motion vector field MVF (see Equation (5)) resulting from mean shift filtering. For a pixel $p = (x, y) \in \Gamma$, let $\Gamma(p)$ denotes its motion value, that is

$$\Gamma(p) = \overline{MV}_i, p \in S_i, P_i = < S_i, \overline{MV}_i, R_i > \tag{6}$$

Let $HP(p)$ and $VP(p)$ denote the projection positions of a pixel p on the horizontal and vertical axis, respectively. For each chosen mode P_i, we define a pair of histogram, H_i and V_i, containing m and n bins respectively, that is

$$\begin{aligned} H_i(x) &= \left|\{p | p \in S_i, HP(p) \text{ is in bin } x\}\right| / |\Gamma|, 0 \le x < m \\ V_i(y) &= \left|\{p | p \in S_i, VP(p) \text{ is in bin } y\}\right| / |\Gamma|, 0 \le y < n \end{aligned} \tag{7}$$

where $|\cdot|$ denotes the set size, $|\Gamma| = n_1 n_2$.

According to the above analysis, we construct the feature vector as follows:

$$\{MV_1, R_1, \ldots, MV_{N_s}, R_{N_s}, H_1(x), V_1(y), \ldots, H_{N_s}(x), V_{N_s}(y)\} \tag{8}$$

where the feature dimension equals $3 * N_s + (n+m) * N_s$.

4. MOTION CHARACTERIZATION

Motion characterization can be considered at two levels: the frame and shot levels. The four tasks mentioned in Section 1 are working at the frame-level. The shot-level characterization can be based on the collective set of frame level labels, e.g. dominant shot motion. Recently a nonparametric probabilistic model [8] was proposed to characterize the dynamic content within a shot from a global point of view. It avoids the frame-level analysis. In this paper, we emphasize the frame-level characterization (i.e. four tasks) using the mean shift based nonparametric motion representation.

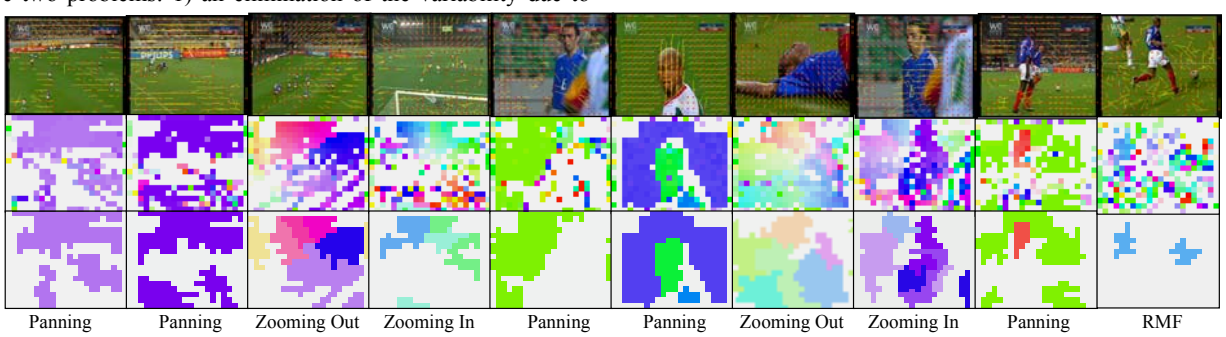

Fig. 3. Examples of mean-shift based MVF representation. First Row: frames overlapped with a MVF, Second Row: representations in MVS space, Third Row: representations resulting from mean shift filtering

5. EXPERIMENTS

Our experiments are performed around those four tasks. All of the MVFs are extracted from P-frames in MPEG-1 compressed video. In order to fairly evaluate the robustness and effectiveness of the nonparametric motion representation, we choose broadcast soccer video. The MVF in soccer video is challenging as a large low-textured field region leads to an incomplete and noisy MVF, and various camera shots exist (Refer to [4]). In our experiments we employ C-Support Vector Classification [11]. The radial basis function (RBF) is selected as kernel function $e^{-g\|x_i - x_j\|^2}$. We use cross-validation to determine the regulation parameter *cost* and the kernel parameter *gamma*. The bin numbers n and m equal 4. The mode number N_s=4. Then the feature dimension equals 44. For each task, we always use ½ as training data and ½ as testing data. Table 1 summarizes our experimental data. Table 2 summarizes the results on testing set.

Table 1. Experimental data (Frames)

Camera Shot	PCM in the absence of POM (846)				POM (125)	RMF (63)	PCM+ POM (62)
	Pan Left/Right	Tilt Up/Down	Zoom In/Out	Diagonal			
Long-shot (LS)	154/144	5/4	44/125	101	0	0	0
Mid-shot (MS)	38/37	3 /4	40/31	50	3	48	0
Close-up (CU)	11/4	5/0	1/20	15	122	15	62

Table 2. Performance of the nonparametric motion characterization

Task		Camera Shot	Gamma	Cost	nSV	Accuracy
POM identification(187/1096)		LS+MS+CU	0.04	64	116	91.07%
PCM identification (908/1096)		LS+MS+CU	0.12	64	143	89.41%
PCM identification in the presence of POM (62/187)		LS+MS+CU	0.04	64	23	92.93%
Recognition of typical forms of video shooting	Zooming (275/1096)	LS	0.05	64	48	95.57%
		MS	0.1	64	27	93.70%
		CU	0.02	64	23	97.67%
		LS+MS+CU	0.07	64	123	91.60%
	Zooming In / Out (96/179)	LS	0.6	64	54	76.92%
		MS	0.4	64	33	84.81%
		LS+MS+CU	0.4	64	90	81.34%
	Panning (415/1096)	LS	0.5	16	102	85.86%
		MS	0.4	16	55	77.31%
		LS+MS+CU	0.5	16	109	84.36%
	Panning Left/Right (212/203)	LS+MS+CU	0.5	16	55	99.25%
	Diagonal (188/1096)	LS	0.4	64	112	82.28%
		LS+MS+CU	0.8	16	231	84.49%
RMF identification (63/1096)		LS+MS+CU	0.1	64	36	98.22%

As listed in Table 1, the Tilt Up/Down is seldom used. Instead the diagonal tilting is widely used to follow actions. We do not consider the Tilt Up/Down for soccer video in our experiments. Note that we summarize data according to camera shots. This is to fairly evaluate our algorithm by executing it in different shot categories as shown in Table 2.

As listed in Table 2, we achieve good performance (89.41% ~ 92.93%) for the identification of POM and PCM. The PCM identification in the presence of POM is higher (3.52%) than the average performance on the whole data. This is due to rich motion patterns and various MVFs in the LS and MS categories. The RMF identification shows quite good accuracy 98.22%. It is practically important for removing non-pattern noisy MVFs. To our surprise, the simplest pattern Panning only achieves 84.36%. After further investigation, we find this is partially due to the ambiguity between panning and diagonal tilting in the case of tilting with a small elevation. Moreover, the POM exhibits the characteristics of panning sometimes. Similarly the performance of Diagonal identification is not high (84.49%). This fact implicitly indicates the introduction of other cues (e.g. color) can improve the performance of Panning and Diagonal. In contrast, we achieve promising results for Zooming (91.60%). It shows the proposed nonparametric motion representation successfully imposes the spatial-range constraints to distinguish zooming from large amounts of distracting patterns. However, the classification between Zooming In/Out is only 81.34%. Through investigating failure MVFs, we find that this is mainly due to the incomplete MVFs induced by large low-texture regions (See Fig.1.). Moreover, it is easy to notice we achieve better results for Zooming identification using tuned parameters (gamma and cost) for each kind of shot category than the whole data. The high performance of Zooming identification is very useful for content-based video indexing. *Readers can contact the authors to request experimental data (around 383MB).*

6. CONCLUSION

We have presented a mean-shift based nonparametric motion characterization scheme. Promising results have been achieved at the frame level. The performance evaluation on the motion fields from various camera shots improves objectivity and fairness to a large extent. Future works include: 1) performing experiments on more extensive video data, 2) performing model comparison among sports video of different types, and 3) investigating the potential of temporal information to enhance the spatial-range motion representation.

7. REFERENCE

[1] M.J. Black and P. Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-smooth Flow Fields," *Computer Vision and Image Understanding* 6(4): 348-365, 1995.

[2] J.-M. Odobez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," *Journal of Visual Communication & Image Representation* 6(4): 348-365, 1995.

[3] P. Bouthemy et al. "A Unified Approach to Shot Change Detection and Camera Motion Characterization," *IEEE Transactions on Circuits and Systems for Video Technology* 9(7): 1030-1044, 1999.

[4] L.-Y. Duan, et al. "A Mid-level Representation Framework for Semantic Sports Video Analysis," In *Proc. of ACM Multimedia'03*, pp. 33-44, 2003.

[5] N.V. Patel and I.K. Sethi, "Video Shot Detection and Characterization for Video Databases," *Pattern Recognition* 30(4): 607-625, 1997.

[6] W. Xiong and J.C.M. Lee, "Efficient Scene Change Detection and Camera Motion Annotation for Video Classification," *Computer Vision and Image Understanding* 71(2): 166-181, 1998.

[7] J.-G. Kim, et al. "Efficient Camera Motion Characterization for MPEG Video Indexing," In *Proc. of ICME'00*, pp.1171-1174, 2000.

[8] R. Fablet, et al., "Nonparametric Motion Characterization Using Causal Probability Models for Video Indexing and Retrieval," *IEEE Transactions on Image Processing* 11(4): 392-407, 2002.

[9] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE Pattern Analysis and Machine Intelligence* 17(8): 790-799, 1995.

[10] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Pattern Analysis and Machine Intelligence* 24(5): 1-18, 2002.

[11] C. Cortes & V. Vapnik, "Support-vector Network," *Machine Learning*, 20:273-291, 1995.