# MOTION DESCRIPTORS FOR SEMANTIC VIDEO INDEXING

Markos Zampoglou
*Department of Applied Informatics, University of Macedonia, Egnatia 156, Thessaloniki, Greece*

Theophilos Papadimitriou
*Department of International Economic Relations and Development, Democritus University of Thrace, Komotini, Greece*

Konstantinos I. Diamantaras
*Department of Informatics, T.E.I. of Thessaloniki, Sindos, Greece*

Keywords: Video Indexing, Motion Information, Bag-of-Motion-Features, Dominant Direction Histogram, PMES.

Abstract: Content-based video indexing is a field of rising interest that has achieved significant progress in the recent years. However, it can be retrospectively observed that, while many powerful spatial descriptors have so far been developed, the potential of motion information for the extraction of semantic meaning has been largely left untapped. As part of our effort to automatically classify the archives of a local TV station, we developed a number of motion descriptors aimed at providing meaningful distinctions between different semantic classes. In this paper, we present two descriptors we have used in our past work, combined with a novel motion descriptor inspired by the highly successful Bag-Of-Features methods. We demonstrate the ability of such descriptors to boost classifier performance compared to the exclusive use of spatial features, and discuss the potential formation of even more efficient descriptors for video motion patterns.

## 1 INTRODUCTION

As capturing and storage devices become cheaper and more widespread and the amount of publicly available multimedia material increases at an ever-greater pace, the need to manage and access this material in an efficient way becomes imperative. Semantic multimedia indexing and retrieval has long been a hopeful research direction, and recently many promising steps have been made. The MPEG-7 standard was an important landmark in this course, by helping standardize widely used descriptors, such as the Color Histogram (Swain and Ballard, 1991), the Edge Direction Histogram (Jain and Vailaya, 1996), and Gabor Texture Descriptors (Bovik, Clark and Geisler, 1990), and since then other powerful descriptors have seen frequent use as well, such as the Color Correlogram (Huang, Kumar, Mitra, Zhu, and Zabih, 1999), Color Moments (Striker and Orengo, 1995), and the Bag-Of-Features descriptors (Zhang, Marszałek, Lazebnik, and Schmid, 2007).

Meanwhile, competitions such as TRECVID and VIDEOCLEF provide us with a measure of success –as well as popularity- for the various descriptors and classifiers proposed. It is within this context that an issue seems to arise: while spatial information is widely used in the form of various descriptors, motion information has faced extremely little use in video content indexing. On the one hand the MPEG-7 motion descriptors have seen limited application: "Camera Motion" requires motion information from the capturing device, which is hardly ever available; "Motion Trajectory" is strictly region- or object-based and is unsuitable for systems that need fixed-length descriptors to characterize scenes; finally, "Motion Activity," despite attempting to capture important information on the motion contents of a shot, is severely hampered by its short length, which is five attributes in all. On the other hand, motion descriptors proposed in TRECVID systems, such as "Motion Moments" (Cao et al, 2006) and the MCG-ICT-CAS "Camera Motion" (Liu et al, 2007) have not as yet gone beyond isolated contributions.



Zampoglou M., Papadimitriou T. and I. Diamantaras K. (2010).
MOTION DESCRIPTORS FOR SEMANTIC VIDEO INDEXING.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 178-184
DOI: 10.5220/0002942601780184

Our aim is to demonstrate the importance of incorporating motion information in a semantic video indexing system, by designing motion descriptors that can significantly boost such a system's performance. To this end, we present a system combining two motion descriptors from our previous research with a novel one, and explore these descriptors' potential in increasing the classification rates of a system otherwise based exclusively on well-established spatial features.

## 2 THE CLASSIFIER SYSTEM

We aim at combining multiple descriptor features for the semantic classification of video shots. The most common way of achieving this is through a 2-level system. In such a system (Figure 1), a number of different classifiers, one per feature, are first trained. From then on, each shot is characterized by a vector of level-1 classifier outputs, with total length equal to the number of features. Consequently the level-2 classifier is trained on these vectors using a different set of video sequences, and its output is the final system output.

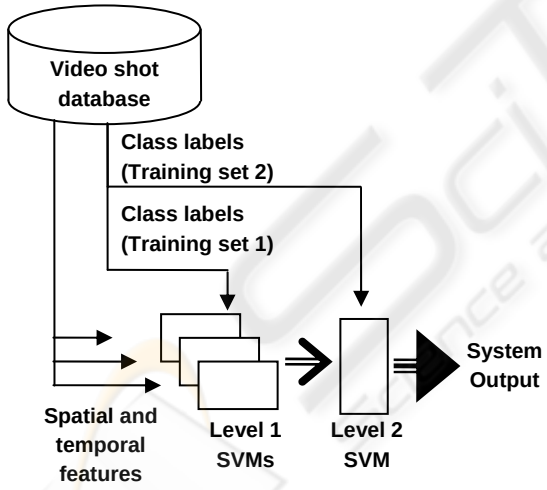


Figure 1: Diagram of the 2-level system.

In our work, we used Support Vector Machines s as classifiers. The classifier parameters used in our experiments are described in Section 4.

## 3 THE MOTION DESCRIPTORS

Three motion descriptors are combined in our system: The Perceived Motion Energy Spectrum (PMES) (Zampoglou, Papadimitriou and Diamantaras, 2007), an extension of the Dominant motion Direction Histogram (DDH) (Zampoglou, Papadimitriou and Diamantaras, 2008) and a novel Bag-of-Motion-Features descriptor (BoMF) inspired by the widely popular Bag-Of-Features approaches. All descriptors presented below are extracted from a shot's sparse motion vector field. Such a field is formed by segmenting each frame into a number of non-overlapping blocks and, consecutively, seeking for each block its best match - according to some similarity criterion- either in the next frame or in one *T* frames in the future.

While it is true that simple block-matching does not necessarily give a very accurate representation of the actual motion events in a shot, it is still a form of information closely related to the actual motion. Furthermore, the fact that it's often already embedded in video data as part of the MPEG compression standard makes it readily available, and, as such, an ideal form of raw motion information.

### 3.1 The Perceived Motion Energy Spectrum

The PMES descriptor is an extension of the work in (Ma and Zhang, 2001) for use in semantic video indexing. This descriptor aims in giving an estimate of foreground motion intensity for the duration of the shot in every part of the frame.

Figure 2: Above: Indicative frames from two shots from our database. Below: The PMES (99 values) descriptor for the same shots.

Having extracted a shot's motion vector field, we begin by estimating the mixture motion energy, i.e. the sum of foreground and background motion

intensity. This is done by calculating the α-trimmed mean magnitude of all motion vectors through time that correspond to every spatial coordinate pair (*i, j*).

$$MixEn_{i,j} = \frac{1}{M - 2\lfloor \alpha M \rfloor} \sum_{m=\lfloor \alpha M \rfloor + 1}^{M-\lfloor \alpha M \rfloor} Mag_{i,j}(m) \qquad (1)$$

In (1), *M* indicates the total number of magnitudes for block (*i, j*) – equal to the number of motion fields in the shot – *α* is the trimming parameter ($0 \le \alpha < 0.5$) and $Mag_{i,j}(m)$ is the magnitude of vector *m*. The mixture energy characterizes each block by the mean (excluding outliers) of all the motion vectors that were assigned to it for the duration of the shot. Thus, it is a measure of the overall motion activity in the corresponding part of the image.

The distinction between foreground and camera motion for the formation of the PMES is based on the assumption of the temporal discontinuity of foreground motions. For camera motions, it is reasonable to assume that they persist, at least in terms of direction, for multiple frames through time. Foreground moving objects, on the other hand, tend to move across the screen passing through blocks. From the perspective of a single block, this means that a foreground object motion causes an abrupt change in the corresponding motion vector's direction that lasts a short period of time. With this in mind, we quantize vector angles into *n* bins and form the angle histogram for every block (*i, j*) through time. By dividing with their sum, we estimate the angle probability distribution for each block (*i, j*). We then proceed to calculate the angle entropies:

$$AngEn_{i,j} = -\sum_{t=1}^{n} p(t) \log p(t) \qquad (2)$$

where $p(t)$ is the probability for angle bin *t*.

The Angle Entropy gives a measure of the variability of vector angles in a block through time, and as such, based on the aforementioned assumption, a measure of the contribution of foreground motion in a block's Mixture Energy. By normalizing Angle Entropies to [0, 1] and multiplying with their corresponding Mixture Energies we end up with a value for each block expressing the amount of foreground motion in that area of the frame for all the duration of the shot (Figure 2).

Ma and Zhang proposed averaging over neighbourhoods, and applied it to content-based retrieval for shots containing specific foreground motion patterns, irrespective of their actual semantic content. In our previous work (Zampoglou et al, 2007), we have shown that the PMES measure can provide a powerful semantic descriptor when combined with Support Vector Machines, for classes whose semantic content is characterised by distinctive motion patterns.

## 3.2 The Dominant Motion Direction Histogram

The DDH is a simple descriptor aiming to express global motion behaviour through time, based on the assumption that, usually, the majority of pixels in a given frame will belong to the background.

To calculate the DDH descriptor, we quantize motion vector directions in a number of bins, and extract the dominant one for each motion field in every video sequence. Obviously, the total number of such dominant values will equal the number of motion fields. By forming the histogram of dominant direction appearances, we can expect to estimate the "history" of the camera motion, provided the dominant vector coincides with the camera motion direction.

In previous approaches, we used four bins: *Horizontal, Vertical, Diagonal* and *Zero Norm*, aiming to keep descriptor length at a minimum. This was a direct result of combining different descriptors by early fusion (feature vector concatenation), where the individual lengths of descriptors had to be as small as possible to reduce the overall dimensionality. Since the system presented here is based on late fusion (2-level SVM), we have no reason to limit classifier length. We thus extend the descriptor to 9 bins, 8 for the vector angles at 45° intervals and one for zero-norm vectors. While further increase of the length (e.g. to 17 or 33 bins) was considered, the idea was finally rejected since the increase in the number of bins actually increased the descriptor's noise sensitivity. Finally, to compensate for different shot lengths, the DDH of each shot is normalized to sum to 1.

## 3.3 The Bag-of-Motion-Features

While the two motion features described above aim at capturing camera and foreground motion in a global sense, we believe that local motion patterns can also provide insights as to the contents of a shot. To explore this possibility, we devised the Bag-of-Motion-Features descriptor (BoMF), inspired by previous work on spatial Bag-of-Features (BoF) methods.

Table 1: The increases in Precision and Recall compared to the exclusive use of spatial features. Set 1: Benchmark set consisting exclusively of spatial features. Sets 2-4: Set 1 plus the DDH, PMES and BoMF respectively. Set 5: All the aforementioned features together.

| | Soccer | | Swimming Pool | | Sea | | Graphics | | Natural | | Urban | | Vehicles | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. |
| Set 1 | 86.82 | 93.35 | 59.43 | 90.63 | 30.20 | 83.33 | 23.21 | 71.00 | 21.26 | 68.33 | 17.98 | 80.78 | 19.80 | 86.00 |
| Set 2 | +2.98 | +0.40 | +7.36 | +1.25 | +4.73 | +0.28 | +3.86 | +2.25 | +2.68 | +4.17 | +1.16 | +1.41 | +0.34 | +0.13 |
| Set 3 | +2.71 | +1.25 | +5.49 | +3.75 | +2.33 | +1.11 | +4.11 | +7.50 | +0.48 | +6.25 | +0.84 | +1.72 | +0.89 | +0.50 |
| Set 4 | +1.07 | +2.27 | +5.87 | +1.87 | +4.87 | +0.84 | +0.48 | +0.00 | +5.51 | +1.67 | +1.67 | +2.35 | +1.05 | +0.88 |
| Set 5 | +5.42 | +2.28 | +17.74 | +3.75 | +8.87 | +2.50 | +12.75 | +8.00 | +7.72 | +6.25 | +3.80 | +2.66 | +2.45 | +1.13 |

The BoF idea comes from the quite successful "Bag-of-Words" methods for text retrieval. These methods completely disregard the structure of a text and treat it as unordered collection of words. Consequently, each text is characterized by the number of instances of each word it contains.

For the Bag-of-Features techniques, certain adaptations had to be made: In place of words, a number of local features are extracted from an image; consecutively, a "dictionary" is formed by clustering all the local features into a small number of classes; finally, for each image, the number of appearances of each "word" is counted, and a histogram is formed for the corresponding probability of appearance. This histogram is then the image descriptor, with length equal to the number of clusters defined in the second step. To our knowledge, such an approach has not as yet been applied to motion. The closest contestant, the "Motion Words" descriptor (Hao, Yoshizawa, Yamasaki, Shinoda, and Furui, 2008), is a traditional local spatial SIFT (Lowe, 2004) descriptor extracted from positions of high inter-frame differences instead of image salient points, and thus can be described as a local spatial descriptor using motion simply for feature detection.

Our BoMF descriptor is formed in a similar manner to the spatial Bag-of-Words methods, but is aimed as a descriptor for local motion information. First, local features are extracted from the motion fields. In spatial bag-of-words methods, it is common practice to extract features from salient points in the image (Mikolajczyk and Schmid, 2005; Quelhas et al, 2005), and, although this practice has been challenged (Jurie and Triggs, 2005), it remains dominant. In our case, however, the sparse motion fields we use as raw information cannot offer the wealth of information usually available for spatial methods. Thus, no salient feature detection is performed, and instead all positions in all fields are considered. Similarly, the local descriptor will have to be small, given the limited overall size of the fields. The local features we use are all possible (overlapping) 2×2 patches of motion vectors from all motion fields, and the descriptor consists of their corresponding four angles.

Consecutively, the local descriptors are clustered using a k-means algorithm in order to form the dictionary. In each step of the clustering algorithm, the mean angle is calculated through the von Mises distribution:

$$\mu = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\} \quad (3)$$

Finally, for each shot, the local patches are assigned to their corresponding clusters and the descriptor vector is formed.

# 4 EXPERIMENTS

Our descriptors were formed as part of an effort to automatically index the archives of the Thessaloniki local TV station "Omega TV". At this stage, experiments are performed on a set of 1074 manually segmented shots of extremely varied content. It was on this material that we performed the descriptor evaluations.

## 4.1 Shot Content and Feature Extraction

The available shots are of size 720×576 with an average length of 178.7 frames. Their content ranges from interviews and newscasts to a wide range of sports shots, graphics, city shots and others. Figure 3 gives an indication of how varied the content within each class is, due to the fact of it comes from a real-world database with practical application in mind.

Figure 3: Indicative frames from four shots of each of the classes tested on the system. First row: Soccer, Swimming Pool. Second Row: Sea, Urban. Third row: Vehicles, Graphics. Bottom row: Natural. Some overlap between classes is visible, e.g. the last "Vehicles" shot shown is also a member of "Urban", while the last "Natural" shot is also a member of the "Sea" class.

Sparse motion vector fields of size 11×9 vectors were extracted from these shots using a hierarchical block-matching algorithm, over a temporal distance $T = 8$ frames, and our motion features were extracted from these fields. The PMES descriptor has length equal to the number of motion vectors, thus 99, and the BoMF features were clustered to length 5000.

Besides the motion descriptors, a number of well-established spatial features were extracted from the shots for comparison. *Color Moments* up to the third order were extracted from the L*a*b* color space, from 8×8 non-overlapping pixel blocks in all channels, resulting in a 576-features long descriptor. The *Color Histogram* was extracted from the HSV color space, with the (*H*, *S*, *V*) channels quantized to 16, 4 and 4 bins respectively, for a total of 256 bins. The *Color Correlogram* was also extracted from the same color space using the same quantization, plus 4 bins for distance, resulting in 1024 bins. The *Edge Direction Histogram* was formed following an application of Canny's edge detection (Canny, 1986) and subsequent quantization of all possible edge directions into 72 bins of 5°, plus a 73rd bin for counting non-edge pixels. Finally, a typical Bag-of-Features descriptor was extracted, where salient points were detected in local extrema of the Difference-of-Gaussians (DoG) scale-space and subsequently represented by a SIFT descriptor. About 10000 local features were extracted from each processed frame, and a dictionary of 5000 visual words was formed by k-means clustering of all features from all images in the database. All these descriptors were extracted from 20 different frames of each shot and averaged in order to capture possible changes in time, with the exception of the Bag-of-Features descriptor which, being computationally more demanding, were extracted from only 5 frames.

Even with this reduction, the computational cost of the Bag-of-SIFT-features is significantly larger than the BoMF descriptor, given that in the latter the clustering is performed on 178.7×99≈17700 local features of length 4 from each shot, in contrast to 5×10000=50000 local features of length 128 for the SIFT features, plus the cost of feature detection.

## 4.2 Experiment Runs and Results

For our descriptors' evaluation, we performed a large number of runs over a range of semantic classes with five different descriptor sets. In all cases, a first-level SVM RBF classifier was trained on each single descriptor, and the results were then fused by a second-level polynomial SVM. *Set 1* consisted only of the five ground-truth spatial descriptors. *Sets 2-4* were similar to the first one, with the addition of one motion descriptor at each time (DDH, PMES and BoMF, respectively). *Set 5* was a fusion of all spatial descriptors with all our motion descriptors. In all cases, the 1st-level training set was 55% of the data and the test set was 25%, leaving 20% for training the 2nd-level committee based on the results of the first. Seven semantic classes were defined, and for each class/descriptor set combination, 40 runs with random training-test sets were performed.

Table 1 demonstrates the relative increase in both measures for Sets 2-5 compared to Set 1. A general improvement can be clearly seen, reaching in certain situations up to almost 18%, although there are significant differences between the various classes. Thus, while for example in the case of *Swimming Pool* shots, containing shots of swimming and water polo, motion information gives a powerful boost, in the case of the *Vehicles* class, which is extremely varied and is not characterized by any specific motion patterns the overall increase is small, while the increases from individual features can be characterised insignificant. Similar explanations can be given for all elements of Table 1. It cannot be denied, however, that the addition of motion information expressed through our descriptors has boosted the classification potential of the overall system.

# 5 CONCLUSIONS AND FUTURE WORK

We have demonstrated the ability of motion descriptors to boost classification performance compared to a system based solely on spatial descriptors. It can be observed, however, that not all descriptors are equally effective for all classes. More in-depth study would help clarify the strengths and limitations of each motion descriptor presented here.

Furthermore, the descriptors presented here could be enhanced in various ways. A dense motion vector field could transform the PMES into a longer and more powerful descriptor which would give a detailed map of foreground activity. It would also allow the local BoMF patches to be expanded beyond size 2×2, thus giving a more detailed description of local motion. Increase in the number of the raw motion vectors would result in an increase in the number of patches, and, as a result salient feature detection could also be performed on the motion patches prior to the local descriptor extraction. In such an approach, the tradeoffs between any increase in performance and the cost of not being able to take direct advantage of MPEG-encoded vector fields should be taken into account.

Finally, rotation invariance should be tried for the BoMF descriptor and compared to its current non-invariant design, while a reduction in the temporal distance during the motion field extraction would lead to a more detailed DDH descriptor, also possibly allowing a narrowing of the angle bins so as to extend of the descriptor length. Furthermore, the DDH descriptor could be made more descriptive by transforming it from a simple histogram of the dominant vector directions to a histogram of actual camera activity, as they can be derived from the motion vector field (Tan, Saur, Kulkarni and Ramadge, 2000).

These further steps in our research would both lead to the formation of more expressive and powerful descriptors for automatic video indexing, and grant us valuable insights as to the ways motion patterns differ between video shots of different semantic content. It is our hope that research on the semantic indexing and retrieval of video shots can eventually reach a point where motion will be considered as important as spatial information for the success of these tasks.

# REFERENCES


Bovik, A. C., Clark, M., Geisler, W. S. (1990). Multichannel Texture Analysis Using Localized Spatial Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12, 55-73.

Canny, J. (1986) A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8, 679–698.

Cao, J., Lan, Y., Li, J., Li, Q., Li, X., Lin, F., Liu, X., Luo, L., Peng, W., Wang, D., Wang, H., Wang, Z., Xiang, Z., Yuan, J., Zheng, W., Zhang, B., Zhang, J., Zhang, L., Zhang, X. (2006). Intelligent Multimedia Group of Tsinghua University at TRECVID 2006. In *Proc. 2006 NIST TREC Video Retrieval Evaluation Workshop*.

Hao, S., Yoshizawa, Y., Yamasaki, K., Shinoda, K., Furui, S. (2008). Tokyo Tech at TRECVID 2008. In *Proc. 2008 NIST TREC Video Retrieval Evaluation Workshop*.

Huang, J., Kumar, S. R., Mitra, M., Zhu, W. J., Zabih, R. (1999). Spatial Color Indexing and Applications. *International Journal of Computer Vision*, 35, 245-268.

Jain, A. K., Vailaya, A. (1996). Image retrieval using color and shape. *Pattern Recognition*, 29, 1233-1244.

Jurie, F., Triggs, B. (2005). Creating Efficient Codebooks for Visual Recognition, *Proc. ICCV '05, 10th IEEE International Conference on Computer Vision*, 1, 604-610 .

Liu, A., Tang, S., Zhang, Y., Song, Y., Li, J., Yang, Z. (2007). TRECVID 2007 High-Level Feature Extraction By MCG-ICT-CAS. In *Proc. 2007 NIST TREC Video Retrieval Evaluation Workshop*.

Lowe, D. G. (2004) Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, 60, 91–110.

Ma, Y. F., Zhang, H. J. (2001). A new perceived motion based shot content representation. In *Proc. ICIP '01, 8th IEEE International Conference on Image Processing*, 3, 426-429.

Mikolajczyk, K., Schmid, C. (2005) A Performance Evaluation of Local Descriptors, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1615-1630

Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T., Van Gool, L. (2005). Modeling scenes with local descriptors and latent aspects. In *Proc. ICCV '05, 10th IEEE International Conference on Computer Vision*, 1, 883-890.

Striker, M. Orengo, M. (1995). Similarity of Color Images. In *Proc. SPIE '95, Storage and Retrieval for Image and Video Databases*, 2420, 381-392.

Swain, M. J., Ballard, D. H. (1991). Color Indexing, *International. Journal of Computer Vision*, 7, 11-32.

Tan, Y.-P., Saur, D. D., Kulkarni, S. R., Ramadge, P. J. (2000). Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, 10, 133-146.

Zampoglou M., Papadimitriou T., Diamantaras, K. I, (2007). Support Vector Machines Content-Based Video Retrieval Based Solely on Motion Information. In *Proc. MLSP '07, 17th IEEE Workshop on Machine Learning for Signal Processing*, 176-180.

Zampoglou, M., Papadimitriou, T., Diamantaras K. I. (2008). From Low-Level Features to Semantic Classes: Spatial and Temporal Descriptors for Video Indexing. *Journal of Signal Processing Systems*, OnlineFirst, doi: 10.1007/s11265-008-0314-3.

Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C. (2007). Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision*, 73, 213-238.