Combining Motion Understanding and Keyframe Image Analysis for Broadcast Video Information Extraction

Ming-yu Chen\textsuperscript{a}, Huan Li\textsuperscript{b}, and Alexander Hauptmann\textsuperscript{a}
\textsuperscript{a}School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213
\textsuperscript{b}School of Computer Science and Engineering, BeiHang University, Beijing, PRC

ABSTRACT

We describe a robust new approach to extract semantic concept information based on explicitly encoding static image appearance features together with motion information. For high-level semantic concept identification detection in broadcast video, we trained multi-modality classifiers which combine the traditional static image features and a new motion feature analysis method (MoSIFT). The experimental result show that the combined features have solid performance for detecting a variety of motion related concepts and provide a large improvement over static image analysis features in video.

Keywords: High-level Feature classification, semantic concepts, video analysis, SIFT, video motion analysis

1. INTRODUCTION

Digital images and motion video have proliferated in the past few years, ranging from ever-growing personal photo collections to professional news and documentary archives. In searching through these archives, digital imagery indexing based on low level image features like color and texture, or manually entered text annotations often fail to meet the user information needs, i.e., there is a persistent semantic gap produced by the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. The video analysis community has long struggled to bridge this semantic gap from successful, low-level feature analysis (color histograms, texture, shape, motion) to semantic content description of video, especially when the video content is as varied as in broadcast news. Early video retrieval systems usually modeled video clips with a set of (low-level) detectable features generated from different modalities. These low-level features like histograms in the HSV, RGB, and YUV color space, Gabor texture or wavelets, and structure through edge direction histograms and edge maps can be accurately and automatically extracted from video. However, because the semantic meaning of the video content cannot be expressed this way, these systems had a very limited success with this approach to video retrieval for semantic queries. Several studies have confirmed the difficulty of addressing information needs with such low-level features.

To fill this “semantic gap”, one approach is to utilize a set of intermediate textual descriptors that can be reliably applied to visual content concepts (e.g. outdoors, faces, animals, etc.). Many researchers have been developing automatic semantic concept classifiers such as those related to people (face, anchor, etc), acoustic (speech, music, significant pause), objects (image blobs, buildings, graphics), location (outdoors/indoors, cityscape, landscape, studio setting), genre (weather, financial, sports) and production (camera motion, blank frames).

The main forum for studying video retrieval aided by semantic concepts has been organized by the National Institute of Standards and Technology in the form of the TRECVID video retrieval evaluations. In 2001, the National Institute of Standards and Technology (NIST) started the TREC Video Track (now referred to as TRECVID) to promote progress in content-based video retrieval via an open, metrics-based evaluation, where the video corpora have ranged from documentaries, advertising films, technical/educational material to multi-lingual broadcast news. As the largest video collections with manual annotations available to the research community, the TRECVID collections have become the standard large-scale testbeds for the task of multimedia retrieval. These evaluations provide a standard collection available to all participants, separated into training and development sets. In 2009, the evaluation data consisted of 100 hours of Dutch TV programs as training data, with another 180 hours as test data. A total of 20 semantic concepts were evaluated.
Success in the semantic concept detection task (also known as high-level feature detection) is measured through precision and recall as the central criteria to evaluate performance. Precision is the fraction of the retrieved documents (or shots in the case of video) that are relevant. Recall is the fraction of relevant documents that are retrieved. NIST also defines another measure of retrieval effectiveness called (non-interpolated) average precision over a set of retrieved documents. Precision $P$ is the number of relevant documents retrieved divided by the total number retrieved. Average precision is defined as:

\[
AP = \frac{\sum_{d=1}^{N} (P(d) \times R(d))}{\text{number of relevant documents in the whole collection}}
\]

where $r$ is the rank, $N$ is the number of retrieved documents, $R(r)$ a binary function on the relevance of a document at the given rank, and $P(r)$ is precision at the given rank.

2. MOSIFT

This section presents our MoSIFT [7] algorithm to detect and describe spatio-temporal interest points. In part-based methods, there are three major steps: detecting interest points, constructing a feature descriptor, and building a classifier. Detecting interest points reduces the whole video from a volume of pixels to compact but descriptive interest points. Therefore, we desire to develop a detection method, which detects a sufficient number of interest points containing the necessary information to recognize a human action. The MoSIFT algorithm detects spatially distinctive interest points with substantial motion. We first apply the well-known SIFT algorithm to find visually distinctive components in the spatial domain and detect spatio-temporal interest points with (temporal) motion constraints. The motion constraint consists of a 'sufficient' amount of optical flow around the distinctive points. Details of our algorithm are described in the following sections.
2.1 MoSIFT interest point detection

Figure 1 demonstrates our MoSIFT algorithm. The algorithm takes a pair of video frames to find spatio-temporal interest points at multiple scales. Two major computations are applied: SIFT point detection and optical flow computation matching the scale of the SIFT points.

SIFT interest points are scale invariant and all scales of an image must be considered. Lowe [1] used a Gaussian function as a scale-space kernel to produce a scale space of the image. The whole scale space is divided into a sequence of octaves and each octave is divided into a sequence of intervals, where each interval is a scaled frame. The number of octaves and intervals is determined by the frame size. The size relationship between two adjacent octaves is in powers of 2. The first interval in the first octave is the original frame. In each octave, the first interval is denoted as $I(x,y)$. We can denote each interval as $L(x,y,k\delta)$. This is done by comparing each pixel in the DoG images to its eight neighbors at the same interval and nine corresponding neighboring pixels in each of the neighboring intervals. The algorithm scans through each octave and interval in the DoG pyramid and detects all possible interest points at different scales.

However, SIFT is designed to detect distinctive interest points in a still image. The candidate points are distinctive in appearance, but they are independent of the motions or actions in video. For example, a cluttered background can produce many interest points unrelated to human actions. Clearly, only interest points with sufficient motion will provide the necessary information for action recognition. The widely used optical flow approach detects the movement of a region by calculating where a region moves in the image space by measuring temporal differences. Compared to video cuboids or volumes, optical flow explicitly captures the magnitude and direction of a motion, instead of implicitly modeling motion through appearance change over time. Our belief is that explicit motion measurement is essential for recognizing actions.

In the interest point detection part of the MoSIFT algorithm, optical flow pyramids are constructed over two Gaussian pyramids. Multiple-scale optical flows are calculated according to the SIFT scales. A local extreme from DoG pyramids can only become an interest point if it has sufficient motion in the optical flow pyramid. We assume that a complicated action can be represented by the combination of a reasonable number of interest points. Therefore, we do not assign strong constraints to spatio-temporal interest points. As long as a candidate interest point contains a minimal amount of movement, the algorithm will extract this point as a MoSIFT interest point. MoSIFT interest points are scale invariant in the spatial domain. However, they are not scale invariant in the temporal domain. Temporal scale invariance could be achieved by calculating optical flow on multiple scales in time. However, we want to select distinctive interest points with sufficient motion such that, ideally, humans could ‘recognize’ the action based on seeing these points, giving us reason to believe that machines should be able to learn a corresponding action model. Therefore, a small motion is sufficient at each interest point rather than imposing a complex motion constraint. Ultimately, this is still an open research topic.

2.2 MoSIFT feature description

In most current work on action recognition, much emphasis is placed on interest point detection and action model learning. However, the feature descriptor is an important component which is only given cursory attention. Most work [8,9,11,12] uses histograms of gradients to describe the appearance of interest volumes or cuboids. Some recent work [13,14] includes histograms of optical flow to boost performance.

Since MoSIFT point detection is based on DoG and optical flow, it is natural that our descriptor should leverage these two features. Instead of combining a complete HoF classifier with a complete HoG classifier, we build a single feature descriptor, which concatenates both HoG and HoF into one vector, which is also known as ‘early fusion’. We believe
appearance and motion information together are the essential components for a classifier. Since an action is only represented by a set of spatio-temporal point descriptors, the descriptor features critically determine the information used in later recognition steps.

It is at times underappreciated that the original SIFT descriptor captures local appearance with an aggregated histogram of gradients from neighboring regions. This gives the SIFT descriptor better tolerance to partial occlusion and deformation. When an interest point is detected, a dominant orientation is calculated and all gradients in the neighborhood are rotated according to the dominant orientation to achieve rotation invariance. The magnitude and direction for the gradient are calculated for every pixel in a region around the interest point in the Gaussian-blurred image \( L \). An orientation histogram with 8 bins is formed for each region, with each bin covering 45 degrees. Each sample in the neighboring window is added to a histogram bin and weighted by its gradient magnitude and its distance from the interest point. Pixels in the neighboring region are normalized into 256 (16x16) elements. Elements are grouped as 16 (4x4) grids around the interest point. Each grid has its own orientation histogram to describe sub-region orientation. This leads to a SIFT feature vector with 128 dimensions (4x4x8 = 128). Each vector is normalized to enhance invariance to changes in illumination. Figure 2 illustrates the SIFT descriptor grid aggregation idea.

MoSIFT adapts the idea of grid aggregation in SIFT to describe motions. Optical flow detects the magnitude and direction of a movement. Thus, optical flow has the same properties as appearance gradients. The same aggregation can be applied to optical flow in the neighborhood of interest points to increase robustness to occlusion and deformation. The main difference to appearance description is in the dominant orientation. Rotation invariance is important to appearance since it provides a standard to measure the similarity of two interest points. In surveillance video, rotation invariance of appearance remains important due to varying view angles and deformations. Since surveillance video is captured by a stationary camera, the direction of movement is actually an important (non-invariant) vector to help recognize an action. Therefore, we omit adjusting for orientation invariance in the MoSIFT motion descriptors. The two aggregated histograms (appearance and optical flow) are combined into the MoSIFT descriptor, which now has 256 dimensions.

3. FEATURES FOR SEMANTIC CONCEPT EXTRACTION

To find the best possible set of features, we extract a number of other image descriptors for each keyframe. Our low level features are derived from static image, motion and audio.
3.1 Low-level features

3.1.1 Grid-based color comments (GCM)
To generate the color moment feature, each image (key-frame) is divided into 5x5 grids, and each grid is described by
the mean, standard deviation, and third root of the skewness of each color channel in the LUV color space. This results in
a 225-dimension (5x5x3x3) color moment feature.

3.1.2 SIFT feature
The local feature of each image is computed from the local key points detected from the image. We use the key points
using the DoG detector and depicted by Scale-invariant feature transform (SIFT) descriptors [1] which describes each
key points by a 128-dimension vector. SIFT features are invariant to image scale and rotation, and are also robust to
changes in illumination, noise, occlusion and minor changes in viewpoint. For each key frame, the number of extracted
key points is different. Therefore, we try to use bag-of-words (BoW) to quantify SIFT feature to a fixed number vector
feature of each key frame. We use K-means clustering to find the conceptual meaningful clusters and each cluster is
treated as a visual word in BoW approach. All the visual words consist of a visual word vocabulary. Then key points in
each key frame are assigned to clusters in the visual vocabulary which are their nearest neighbors. In the end, each key
frame is presented by a visual word histogram feature. The performance of BoW in high-level feature detection in large-
scale multimedia corpus is subject to several aspects, such as the size of the visual word vocabulary, visual word
weighting scheme, etc. We discuss these factors next.

- Size of vocabulary
While text vocabulary size is relatively fixed in text information retrieval, the size of a visual words vocabulary is
decided by the number of clusters generated by clustering process. Choosing vocabulary size is a trade-off between
discrimination and generalization. A small vocabulary is less discriminative since two keypoints may be assigned into
one cluster even if they are not similar to each other. On the other hand, a large vocabulary may lack of generalizability
since similar keypoints may be assigned to different clusters. And it also increases the cost associated with clustering
keypoints, assigning each keypoint to the cluster and running supervised learning with high dimension features. Our
previous work shows using a moderate visual word vocabulary size lead a better performance. So we cluster the key
points into 1000 clusters and at the same time we use distributed processing to reduce the computation time.

- Soft cluster boundary
Term weighting is a critical problem in text information retrieval. Term frequency (tf) and inverse document frequency
(idf) are mostly used with BOW features. Essentially, this term weighting scheme assigns a key point to its nearest
neighbor cluster without considering the relationship between this keypoint to other nearby clusters. This kind of
assignments is called hard boundary which ignores the information of other nearest neighbors, e.g., the second
nearest cluster.

In our experiment, we consider N nearest neighbors of one key point and assign different weights to clusters according to
their distance rank. For each key point in an image, we select N (N=4) nearest neighbor clusters for it. These N nearest
neighbor clusters are then assigned weights with their inverse rank value. The final weight of each cluster is the sum of
inverse rank values calculated from all the keypoints in an image belong to it. Suppose we have a visual vocabulary of K
visual words, we have a K-dimension feature vector \( v = \{v_1, v_2, \ldots, v_K\} \) for each image where each term represents the
weight of \( i \)th visual word in the image,

\[
v_k = \sum_{i=1}^{N} \sum_{j=1}^{B_i} \frac{1}{2^{i-1}}
\]

where \( B_i \) represents the number of key points whose \( i \)th nearest neighbor is \( k \). In particular, normalization factor is
significant since different images may have different number of key points. Even among images of the same size, the
number of keypoints varies according to the complexity of the image content. Normalization eliminates such difference.
So we normalize our term weight such as
\[ V_k' = \frac{V_k}{\sum_{m=1}^{M} \sum_{i=1}^{N} \frac{1}{2^{i-1}}} \]  

(4)

where M is the number of key points in an image. We apply both hard boundary and soft boundary to calculate the term weight. The result is shown in Table 1.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Descriptor</th>
<th>Construction</th>
<th>Vocabulary size</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard boundary</td>
<td>SIFT</td>
<td>K-MEANS</td>
<td>1000</td>
<td>0.068</td>
</tr>
<tr>
<td>Soft boundary</td>
<td>SIFT</td>
<td>K-MEANS</td>
<td>1000</td>
<td>0.089</td>
</tr>
</tbody>
</table>

3.1.3 MOSIFT feature
MoSIFT is described in Section 2. It is a descriptor which explicitly describes appearance and motion for a region which contains abundant information. MoSIFT is represented as a bag-of-word feature too for each shot. We also apply soft boundary to form the bag-of-word feature.

3.1.4 Texture feature
The texture features are obtained from the convolution of the image pixels with Gabor wavelet filters. We compute it in 7*7 image grids. In each grid we use the mean and variance of twelve oriented energy filters aligned in 30-degree intervals.

3.1.5 Face feature
Schneiderman’s face detector [15] is used to detect the faces from the video key frames with a confidence score, pose, scale and location for each detected face in the keyframe.

3.1.6 Audio feature
Mel-frequency cepstral coefficients (MFCCs) [16] are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression. We create a 20-dimensional MFCCs feature for each key frame based on this.

3.2 Kernel-based learning
Similar to previous years, we evaluate a set of SVMs with different kernels using different features and model parameters for each high-level feature. For this, we use the LIBSVM implementation [2] of SVM with probabilistic output [3, 4].

- Cross-validation
The parameters of SVM are well known to have a significant influence on performance. For each parameter combination, we compute its performance in TRECVID 2007 development data using a 2-fold cross-validation to prevent over-fitting. Performance is measured by average precision (AP). We select the parameter combination that yields the best performance to train SVM models for each high-level feature using TRECVID 2007 development data and test data. This results in models which we will then use for late fusion.
SVM kernels

We divide all these local features and global features into two categories: histogram features and non-histogram features. Histogram features are features such as SIFT and MOSIFT features which are represented by frequencies of the visual words in an image. Non-histogram features are features such as the grid-based color moments feature which is a concatenation feature over all grids. For histogram features, we apply a $x^2$ kernel in SVM because it has been shown to be better for calculating histogram distances [5]. The $x^2$ kernel is defined as

$$K(x_i, x_j) = \exp \left( -\frac{1}{A} D(x_i, x_j) \right)$$

where $A$ is a scaling parameter that can be determined through cross-validation. $D(x_i, x_j)$ is the $x^2$ distance defined as:

$$D(x_i, x_j) = \frac{1}{2} \sum_{j=1}^{m} \left( \frac{u_i - w_j}{u_i + w_j} \right)^2$$

with $x_i = (u_1, ..., u_m)$ and $x_j = (w_1, ..., w_m)$.

Radian basis kernel (RBF) in SVM is applied for non-histogram features.

3.3 Late Fusion

It was shown by Snoek [6] that late fusion frequently has better performance for most high-level features than early fusion. Therefore, we use 2 kinds of late fusion strategies to combine the prediction results from different low-level features. One strategy is named Meta fusion which takes the component probability output as input and outputs an overall prediction. The other one is named Borda-rank which uses the value of the inverse rank instead of the probability output as input. For both strategies we use SVM to train the final prediction model. TRECVID2007 test data are used for a 2-fold cross validation with an RBF kernel, from which select the one with the best performance parameters. Since different low-level features have different prediction performance, we select different combinations of these features for late fusion. Meta fusion of the SIFT feature, MOSIFT feature, color feature and the texture feature exhibits the highest MAP which is 0.139 in our test evaluation using TRECVID2007 test data.

4. EXPERIMENTAL RESULTS

In order to evaluate the benefits and drawbacks of different features and combination strategies, we submitted a total of 6 runs to NIST for evaluation. The runs can be described as follows:

- **A_CMU1_1**: SIFT feature alone, trained with $x^2$ kernel for each high-level feature.
- **A_CMU2_2**: MOSIFT feature alone, trained with $x^2$ kernel for each high-level feature.
- **A_CMU3_3**: Meta fusion of A_CMU1_1 and A_CMU2_2 for each high-level feature.
- **A_CMU4_4**: Meta fusion of A_CMU3_3 with Support Vector Machine (SVM) classification results of color feature and texture feature.
- **A_CMU5_5**: Meta fusion of A_CMU4_4 with SVM classification results of audio feature and face feature.
- **A_CMU6_6**: Select the best performing classifier (on the training data) for each high-level feature by using different feature combinations and late fusion strategies.

Figure 3 shows an overview of all our submitted high-level feature runs as evaluated by NIST. The average performance for each run is shown in the last column. The A_CMU2_2 run shows the best performance with a MAP of 0.112 in our 6 runs. The MAP of the worst run A_CMU1_1 scored much lower than the others. For some high-level features such as People-dancing, Hand, Airplane_flying, Person-playing-soccer, our performance have outperformed almost all other approaches also when compared against other systems. But for some other features such as Chair, Classroom, and Singing, our performance is far from satisfactory.
5. CONCLUSIONS

The low average scores shows that this task is quite challenging, even though our approaches scored significantly above the median compared to other submitted approaches at TRECVID 2009.

Our experiments with the soft boundary clustering showed great success, hard decisions seem to be quite detrimental to a good classification result. This year we used a soft boundary for assignment of key points to clusters and the performance increased of 31.4% compared to the traditional hard-boundary assignment of each point to exact one cluster.

The motion-based MOSIFT feature resulted in excellent performance on a majority of the features. This was a little surprising in that the SIFT features contained half the same feature vector, yet performed poorly. This can only be attributed to the motion component of the MoSIFT classifier. The A_CMU2_2 run achieved the highest MAP and had the best performance for 8 high-level features in our submitted 6 runs. It solely used the MOSIFT feature. This was not combined with any other low-level feature set. Especially concepts that involved motion, such as playing soccer, moving hands or people dancing seem to clearly benefit from the motion based approach.

By pre-computing the kernel matrix for SVM, we obtained significantly reduced computation time. We pre-compute kernel matrix which reduces SVM computation time, which is especially important when many iterations are required to tune different parameters. A distributed, parallel computation of k-means was implemented, which further reduced clustering computation time.

6. ACKNOWLEDGEMENTS

This work was supported by the Nation Science Foundation under Grant No. IIS-0205219 and Grant No. IIS-0705491. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
7. REFERENCES