# Words-of-Interest Selection based on Temporal Motion Coherence for Video Retrieval

Lei Wang[1], Dawei Song[1], Eyad Elyan[1]
[1]School of Computing, Robert Gordon University, United Kingdom
{l.wang4, d.song, e.elyan}@rgu.ac.uk

## ABSTRACT
The "Bag of Visual Words" (BoW) framework has been widely used in query-by-example video retrieval to model the visual content by a set of quantized local feature descriptors. In this paper, we propose a novel technique to enhance BoW by the selection of Word-of-Interest (WoI) that utilizes the quantified temporal motion coherence of the visual words between the adjacent frames in the query example. Experiments carried out using TRECVID datasets show that our technique improves the retrieval performance of the classical BoW-based approach.

## Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information search and Retrieval – *Multimedia Information Retrieval*

## General Terms
Algorithms, Performance, Theory

## Keywords
Bag of visual Words, Words-of-Interest, Video Retrieval, Temporal Motion Coherence

## 1. INTRODUCTION
Content-based video retrieval is an important and challenging task in multimedia information retrieval. We are especially interested in the case of query-by-example, i.e. copyright infringement detection, to find videos relevant to an example video issued by the user.

Inspired by the success in language modeling and text retrieval, visual content representation based on Bag of Visual Words (BoW) has been wildly investigated and has shown some impressive results in content-based video retrieval [2], image classification and object recognition. The BoW model is based on the assumption that all the visual words are generated independently, and the spatial-temporal relationships between the visual words are ignored to achieve better computational and representational efficiency. In the context of video understanding and retrieval, however, a single visual word contains only small amount of information, and the temporal relationships among the different words are sometimes critical to represent the visual information on how the objects behave.

Recently, several techniques have been proposed to address these limitations. Cao et al. [1] proposed a Spatial Coherence Latent Topic Model (Spatial-LTM) that groups visual words with respect to latent topics in order to improve the representative power. Liu and Chen [6] argued that spatial and temporal information could be used to extract the object of interest for video retrieval.

In this paper, we aim to discover the temporal relations between visual words and explore the possibility to enhance the BoW representation for video retrieval. Inspired by the method proposed by Wang et al. [5] which classifies the relative motion of visual words to represent the temporal patterns in a video, we propose to utilize relative motion to model the temporal relation of visual words.

According to [3], a subset of visual words is more descriptive for certain catalogue of objects than others. In context of query-by-example video retrieval, we believe that some selected visual words, called Words-of-Interest (WoI), are more related to the user's interest than others. The WoI should be given higher weights for similarity match between query and video data.

In this paper, we hypothesize that the WoI would appear and move in a relatively coherent manner in a video, while non-WoI occurs more singularly and randomly. Here, the *temporal motion coherence* can be defined as the degree to which a visual word moves coherently with other words on the temporally aligned frames in the query video.

The reminder of this paper is organized as follows. In section 2, our temporal motion coherence based WoI generation algorithms and the application in video retrieval are presented. The experimental results are discussed in Section 3. Finally, Section 4 concludes the paper and highlights some future research directions.

## 2. WoI Selection Based on Temporal Motion Coherence
In the BoW model, a number of local feature descriptors are automatically extracted, e.g., using the Speed Up Robust Feature algorithm (SURF) [4]. A visual vocabulary $\text{Voc} = \{w_j\}, j = 1 \ldots \text{N}_\text{w}$ is generated by clustering a large amount of descriptors, e.g., by the K-means algorithm. Then a video $V$ is represented as a set of key frames $\{f_c\}$ (sampling 1 key-frame per ten frames), where each $f_c$ is a weighted vector of visual words, and the weight of each visual word is often its Term Frequency (TF) in the frame. Similarly, a given query example is also represented as a set of key frames $\{f_q\}$. The distance $|f_c - f_q|$ is computed for frame level similarity match.

### 2.1 Quantifying Temporal Motion Coherence
In this section, the temporal motion coherence defined in Section 1 is quantified by the relative motion between the visual words pair in the given query.

The motion of instances of a visual word can be extracted between every two neighboring frames by an algorithm based on $L_2$ norm [8] in the next frame. Specifically, for each instance $p$ of a visual word $w$ that can be successfully tracked, its motion $\boldsymbol{m_p} = (m_x, m_y)$ is calculated as a vector, where $m_x$ and $m_y$ identify the vertical and horizontal motion respectively. Each instance $p$ is associated with a motion vector $\boldsymbol{m_p}$ and a visual word $w \in \text{Voc}$.

To evaluate how coherently a pair of visual words moves in neighboring frames, we calculate a relative motion vector for them. The relative motion vector $\overline{m}_{i,j}$ of the visual words pair $w_i$ and $w_j$ in the query video is formulated as follows:


*SIGIR'11*, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

$$\bar{m}_{i,j} = \frac{1}{N_i * N_j} \sum_{k=1}^{N_i} \sum_{n=1}^{N_j} (\boldsymbol{m_{p,i}}^k - \boldsymbol{m_{p,j}}^n) \quad (1)$$

where $\boldsymbol{m_{p,i}}^k$ and $\boldsymbol{m_{p,j}}^n$ are the motion vectors of the $k^{\text{th}}$ instance of $w_i$ and the $n^{\text{th}}$ instance of $w_j$, and $N_i$ and $N_j$ indicate the number of instances of $w_i$ and $w_j$ respectively.

The smaller the relative motion vector, the higher the degree of motion coherence of the visual words pair. In order to quantify the temporal coherence of a visual word, a simplification is proposed. The relative motion vectors of a visual word with all other words are averaged as:

$$\boldsymbol{r_i} = \sum_{j=1}^{N_w} |\bar{\boldsymbol{m}}_{i,j}| * \boldsymbol{p(w_j)} \ , \boldsymbol{p(w_j)} = \frac{\boldsymbol{tf_j}}{\boldsymbol{N_q}} \quad (2)$$

where $p(w_j)$ is the conditional probability that $w_j$ occurs in the query $q$, $tf_j$ is the term frequency of the visual word $w_j$ and $N_q$ is the number of instances of the visual words in $q$.

## 2.2 Selecting the WoI

Based on our hypothesis in Section 1, the WoI with high motion coherence can be selected when the corresponding $\mathrm{r}_i$ values are lower than an empirical threshold. However, a pre-fixed empirical threshold may not be suitable for every query. To tackle this problem, the EM algorithm [7] is used to classify the visual words based on temporal motion coherence adaptively.

More formally, each visual word $w_i$ is associated with a hidden variable $z \in \{z_+, z_-\}$. Here, $z_+$ denotes that $w_i$ is WoI, and $z_-$ denotes that $w_i$ is out of user interest. Naturally, $p(z_+)$ represents the probability of a visual word belonging to WoI. The motion coherence of visual words is modeled as $p(r|z)$. We assume that $p(r|z_+)$ and $p(r|z_-)$ are both Gaussian distributions. From these definitions, the joint distribution of $p(z,r)$ is defined as $p(z,r) \equiv p(z)p(r|z)$, and we simplify the problem by assuming $z$ and $r$ are independent variables. All distributions are unknown yet, and the parameters should be estimated using the EM algorithm.

The steps of the EM algorithm for estimating the unknown distribution is given as follows:

**Table 1: EM algorithm for WoI selection**

**E-step:**

$p(\boldsymbol{z}|r) = \mathrm{c}_1 p(\boldsymbol{z}) p(r|\boldsymbol{z})$

$\mathrm{E}_{p(z|r)}[\log\, p(r|\Phi)] = \prod_i \sum_i p(\boldsymbol{z}_i) p(r_j|\boldsymbol{z}_i)$

**M-step:**

$\Phi_{new} = argmax_{\Phi}(\mathrm{E}_{p(z|r)}[\log\, p(r|\Phi)])$

where $\Phi$ is a set of parameters to be estimated, $\mathrm{c}_1$ is the nominalization factor to guarantee that the sum of $p(\boldsymbol{z}|r)$ equals 1. The estimated $p(r|z_+)$ identifies the location of WoI in the temporal motion coherence space. We choose the visual words which are located within the standard deviation of the Gaussian distribution $p(r|z_+)$ as WoI:$WoI = \{w_j\}, j = id_1 \ldots id_k \ldots id_{\mathrm{N_{woi}}}$.

The frame level similarity based on the WoI is measured by the distance:

$$\bar{\mathbf{D}} = |\boldsymbol{f_c} - \boldsymbol{f_q}| + \boldsymbol{\alpha_{woi}} \times |\acute{\boldsymbol{f}}_{\boldsymbol{c}} - \acute{\boldsymbol{f}}_{\boldsymbol{q}}| \quad (3)$$

where $\acute{f}_c$ represents the term frequency of WoI in the key frame of the video $c$ and $\acute{f}_q$ represents the key frame of the query, and $\alpha_{\mathrm{woi}}$ is an empirically selected weighting factor.

Finally, the videos with higher number of key frames similar to the key frames from the query are ranked higher in the results.

# 3. Experimental Results

To evaluate our temporal motion coherence based method, we select a dataset from TRECVID 2002, which consists of 3000 videos and 6 query topics. The evaluation is based on common criteria used in the information retrieval community: Precision, Recall and Mean Average Precision (MAP).

As shown in Figure 1, the WoI enhanced BoW model (denoted WoI) outperforms the classical BoW model for the query-by-example video retrieval task. Most points of the WoI curve are above the curve of classical BoW.

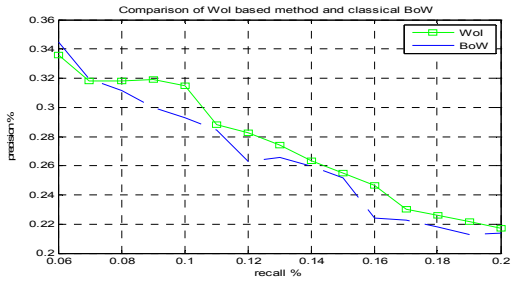


**Figure 1: the performance comparison between WoI enhanced BoW model and classical BoW**

Moreover, for 5 out of 6 topics, the WoI based method outperforms the classical BoW in term of Average Precision. The MAP (over all the 6 test topics) of WoI outperforms BoW by 2.3%, which is statistically significant (*P-value* = 0.02188).

This result demonstrates the feasibility of modeling the interest of a user by temporal motion coherence. We believe further improvement could be achieved by better quantification methods, as currently the approach in Section 2.1 may have over-simplified the motion coherence of visual words.

# 4. Conclusions and Future Work

In this paper, we propose a technique for WoI selection based on the quantified temporal motion coherence in the query video example. The experimental results show that the selection is effective and the generated WoI could improve query-by-example video retrieval based on classical BoW model.

Our future work will be focused on more descriptive user interest representation model which will incorporate spatial-temporal information.

# 5. References


[1] L. Cao and L. Fei-Fei. "Spatially coherent latent topic model for concurrent object segmentation and classification" Vol. 2. pp. 674-679 ICCV(2007)

[2] J. Sivic and A. Zisserman, "Efficient visual search for objects in videos," Proceedings of the IEEE, vol. 96, no. 4, pp. 548-566, 2008

[3] S. Zhang, Q. Tian, G. Hua, Q. Huang, S. Li, 'Descriptive Visual Words and Visual Phrases for Image Applications', in ACM MM, pp. 75-84 (2009)

[4] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, "SURF: Speeded Up Robust Features", CVIU, pp. 404-417 (2008)

[5] F. Wang, Y. Jiang, and C. Ngo. "Video event detection using motion relativity and visual relatedness", ACM MM, pp. 239-248 (2008).

[6] D. Liu and T. Chen. "Video retrieval based on object discovery". pp. 397-404 CVIU (2009)

[7] A. P. Dempster , N. M. Laird , D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm", Journal Of The Royal Statistical Society, Series B Vol. 39, No. 1. (1977), pp. 1-38 (1977)

[8] B. D. Lucas and T. Kanade, "An Interative Image Registration Technique with an Application to Stero Vision", IJCAI, Vol. 2, pp. 674-679 (1981)