# TEMPORALLY CONSISTENT GAUSSIAN RANDOM FIELD FOR VIDEO SEMANTIC ANALYSIS

*Jinhui Tang* †, *Xian-Sheng Hua* ‡, *Tao Mei* ‡, *Guo-Jun Qi* †, *Shipeng Li* ‡, *Xiuqing Wu* †

† Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei, 230027 China
‡ Microsoft Research Asia, Beijing, 100080 China

## ABSTRACT

As a major family of semi-supervised learning, graph based semi-supervised learning methods have attracted lots of interests in the machine learning community as well as many application areas recently. However, for the application of video semantic annotation, these methods only consider the relations among samples in the feature space and neglect an intrinsic property of video data: the temporally adjacent video segments (e.g., shots) usually have similar semantic concept. In this paper, we adapt this temporal consistency property of video data into graph based semi-supervised learning and propose a novel method named Temporally Consistent Gaussian Random Field (TCGRF) to improve the annotation results. Experiments conducted on the TRECVID data set have demonstrated its effectiveness.

***Index Terms***— video annotation, temporal consistency, graph based method

## 1. INTRODUCTION

With the decreased cost of storage devices, high transmission rates and improved compression techniques, digital videos are prevailing at an ever increasing rate. The demand for solutions to manage large scale video database is increasing tremendously. It is a common theme to develop the automatic analysis techniques for deriving metadata for describing information in the content at both syntactic and semantic levels. With the help of these metadata, the tools and systems for video retrieval, summarization, delivery and manipulation can be created effectively.

Automatic semantic annotation (or we may call it concept detection or high-level feature extraction) of video or video segments is an elementary step for obtaining these metadata. For general automatic video annotation methods, statistical models are built from manually pre-labeled samples, and then the labels of unlabeled samples can be estimated using these models. However, the major obstacle of this process is that, the labeled data is limited, so that the distribution of the labeled data can not well represent the distribution of the entire data set (include labeled and unlabeled). This kind of insufficiency of labeled data usually leads to inaccurate annotation results.

Semi-supervised learning techniques [5], which attempt to learn from both labeled and unlabeled data, are promising to solve the above problem. As a major family of semi-supervised learning, graph based methods have attracted more and more researchers' attention recently [13][14][12][4]. They have been successfully applied in text categorization [3] and image annotation [8]. Meanwhile, some graph based methods have also been proposed for video annotation. In [11], a manifold ranking method based on feature selection is proposed for video concept detection. An anisotropic manifold ranking method is proposed in [9] for video semantic annotation, where the authors analyze the graph-based semi-supervised learning methods from the view of PDE based diffusion.

However, existing graph based methods neglect an important and intrinsic property of video data called temporal consistency, that is, with high possibility temporally adjacent video segments (e.g., shots) will be related to a same semantic concept. For example, if a shot in the video matches the concept *sports*, most likely a few shots previous and next to this shot are also about sports. The authors have shown in [10] that this property is an important clue for semantic video annotation. In this paper, we adapt the temporal consistency assumption into graph based semi-supervised learning by combining the temporal consistency and feature space similarity, and propose a novel method named Temporally Consistent Gaussian Random Field (TCGRF) for video annotation. Experiments conducted on the TRECVID [1] data set show that this approach significantly improves the annotation performance compared with existing Gaussian Random Field (GRF) [14] based methods.

The rest of this paper is organized as follows. In Section 2, we detail the algorithm of TCGRF; and Section 3 presents the regularization framework for TCGRF; Experiments are introduced in Section 4, followed by the conclusion remarks in Section 5.

This work was performed when the first and fourth authors were research interns at Microsoft Research Asia.

## 2. TEMPORALLY CONSISTENT GAUSSIAN RANDOM FIELD

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of $n$ samples (i.e., video shots in our application) in $R^m$ (feature space of $m$ dimensions). The first $l$ points are labeled as $\mathbf{y} = [y_1, y_2, \ldots, y_l]^T$ with $y_i \in \{0, 1\}$ $(1 \leq i \leq l)$ and the remaining points $x_u$ $(l+1 \leq u \leq n)$ are unlabeled. Consider a connected undirected graph $G = (V, E)$ with node set $V = L \bigcup U$ corresponding to the $n$ data points, where the node set $L = \{1, \ldots, l\}$ contains labeled points and node set $U = \{l+1, \ldots, l+u\}$ are unlabeled ones. The edges $E$ are weighted by the $n \times n$ affinity matrix $W$ with entries

$$w_{ij} = exp\{-\frac{\|x_i - x_j\|^2}{2\sigma^2}\} \tag{1}$$

when $j \neq i$ and $w_{ii} = 0$. This is from a basic assumption in graph based semi-supervised learning: nearby points are likely to have the same label.

Let vector $\mathbf{f} = [f_1, f_2, \ldots, f_l, f_{l+1}, \ldots, f_n]^T = [\mathbf{f}_L^T, \mathbf{f}_U^T]^T$ denote the predicted labels of $X$, where the superscript $T$ denotes transpose. Another assumption in graph based semi-supervised learning method is the labels should vary smoothly in the feature space. So GRF method [14] proposed to minimize the energy function

$$Q(f) = \frac{1}{2} \sum_{1 \leq i,j \leq n} w_{ij}(f_i - f_j)^2 \tag{2}$$

subject to $\mathbf{f}_L = \mathbf{y}$. It has been shown in [14] that the minimum energy function $f = argmin_{\mathbf{f}_L = \mathbf{y}} Q(f)$ is harmonic. Therefore, it satisfies $\Delta f = 0$ on the unlabeled data points $U$, and is equal to $\mathbf{y}$ on the labeled data points $L$. Here $\Delta$ is the *combinatorial Laplacian* [6] with matrix form $\Delta = D - W$, where $W$ is the affinity matrix and the $D = diag(d_i)$ is the diagnal matrix with entries $d_i = \sum_{j=1}^{n} w_{ij}$. The harmonic property results in that the value of $f$ at each unlabeled point is the weighted average of $f$ at other points:

$$f_i = \frac{1}{d_i} \sum_{j=1}^{n} w_{ij} f_j \quad i \in U. \tag{3}$$

Although GRF method has been well applied in text categorization and image annotation, for video data, it only considers the relations among samples in the feature space and neglect the temporal consistency. We believe that temporal consistency provides valuable contextual clues to video semantic annotation.

According to the temporal consistency, nearby points over the temporal order may have similar labels. We define a measurement of the probability that two samples have the same label in temporal order index $i$ and $j$:

$$h_{ij} = exp\{-\frac{(i-j)^2}{2\sigma_t^2}\}, \tag{4}$$

where $\sigma_t$ is a scale parameter over the temporal order.

Define the following energy function:

$$R(f) = \frac{1}{2} \sum_{1 \leq i,j \leq n} h_{ij}(f_i - f_j)^2, \tag{5}$$

the low energy corresponds to a slowly varying function over the temporal order. Minimizing $R(f)$ subject to $\mathbf{f}_L = \mathbf{y}$ also results in harmonic function $f$:

$$f_i = \frac{1}{d_i^{'}} \sum_{j=1}^{n} h_{ij} f_j, i \in U \tag{6}$$

where $d_i^{'} = \sum_{j=1}^{n} h_{ij}$.

Here we adapt the temporally consistency into GRF by combining the temporal order adjacency and the feature space similarity. Then we have

$$\begin{aligned} f_i &= (1-\alpha)\frac{1}{d_i} \sum_{j=1}^{n} w_{ij} f_j + \alpha \frac{1}{d_i^{'}} \sum_{j=1}^{n} h_{ij} f_j, i \in U \\ &= \sum_{j=1}^{n} ((1-\alpha)\frac{w_{ij}}{d_i} + \alpha \frac{h_{ij}}{d_i^{'}}) f_j \\ &= \sum_{j=1}^{n} p_{ij} f_j \end{aligned} \tag{7}$$

where $p_{ij} = (1-\alpha)\frac{w_{ij}}{d_i} + \alpha \frac{h_{ij}}{d_i^{'}}$, and $\alpha$ controls the effectiveness of the temporal consistency among the two effects. Representing (7) in matrix form, we have

$$\mathbf{f} = ((1-\alpha)D^{-1}W + \alpha D'^{-1}H)\mathbf{f} = P\mathbf{f} \tag{8}$$

subject to $\mathbf{f}_L = \mathbf{y}$, where $P = (1-\alpha)D^{-1}W + \alpha D'^{-1}H$, $D = diag(d_i)$, $D' = diag(d_i^{'})$. Split the matrix $P$ after the $l$-th row and $l$-th column

$$P = \begin{bmatrix} P_{LL} & P_{LU} \\ P_{UL} & P_{UU} \end{bmatrix}. \tag{9}$$

Substitute the $P$ in (8) with Eq.(9), substitute the $\mathbf{f}$ with $[\mathbf{f}_L^T, \mathbf{f}_U^T]^T$, and solve the obtained equations we will obtain the optimal labels for the unlabeled samples in matrix form as follows

$$\mathbf{f}_U^* = (I - P_{UU})^{-1} P_{UL} \mathbf{y}, \tag{10}$$

where $I$ is the identity matrix. From Eq.(10), each sample will be assigned a real-value score indicating the degree of belonging to a specific concept.

Consequently, the algorithm of TCGRF is summarized in **Algorithm 1**.

## 3. REGULARIZATION FRAMEWORK

In this section we will show that our method could also be obtained from a regularization framework. The cost function

**Algorithm 1** TCGRF

1: Calculate the affinity matrices $W$ and $H$ over fea space and temporal order respectively;
2: Construct matrix $P = (1-\alpha)D^{-1}W + \alpha D'^{-1}H$ which $D$ is a diagonal matrix with its $(i,i)$-elen equals to the sum of the $i$-th row of $W$, and $D'$ is a agonal matrix with its $(i,i)$-element equals to the sur the $i$-th row of $H$;
3: Split the matrix $P$ into $P_{LL}, P_{LU}, P_{UL}$ and $P_{UU}$ acc ing to (9);
4: Predict the real-value labels for unlabeled samples $\mathbf{f}_U^* = (I - P_{UU})^{-1}P_{UL}\mathbf{y}$.

associated with $f$ is defined as:

$$E(f) = \frac{1-\alpha}{2}\sum_{1\le i,j\le n}\frac{w_{ij}}{d_i}(f_i - f_j)^2 + \frac{\alpha}{2}\sum_{1\le i,j\le n}\frac{h_{ij}}{d'_i}(f_i - f_j)^2 + \infty\sum_{1\le i\le l}(f_i - f_j)^2 \quad (11)$$

Re-write it into a matrix form:

$$E(\mathbf{f}) = (1-\alpha)\mathbf{f}^T(I - D^{-1}W)\mathbf{f} + \alpha\mathbf{f}^T(I - D'^{-1}H)\mathbf{f} + \infty(\mathbf{f}_L - \mathbf{y})^T(\mathbf{f}_L - \mathbf{y}) \quad (12)$$

Minimizing the above cost will result in the optimal $\mathbf{f}^*$:

$$\mathbf{f}^* = argmin_{\mathbf{f}}E(\mathbf{f}) \quad (13)$$

Differentiating $E(\mathbf{f})$ with respect to $\mathbf{f}$, we will obtain:

$$(1-\alpha)(I - D^{-1}W)\mathbf{f} + \alpha(I - D'^{-1}H)\mathbf{f} + \infty(\mathbf{f}_L - \mathbf{y}) = 0,$$

which can be transformed to:

$$\mathbf{f} = (1-\alpha)D^{-1}W\mathbf{f} + \alpha D'^{-1}H\mathbf{f}. \quad s.t. \quad \mathbf{f}_L = \mathbf{y} \quad (14)$$

This is the same as we obtained in formula (8). It demonstrates we can obtain the same result of TCGRF from this regularization framework.

## 4. EXPERIMENTAL RESULTS

In the following experiments, we use the video data set of the TRECVID 2005 corpus, which is consisted of 170 hours of TV news videos from 13 different programs in English, Arabic and Chinese. After automatic shot boundary detection, the development (DEV) set contains 43907 shots, and the evaluation (EVAL) set contains 45766 shots. Some shots are further segmented into sub-shots, and there are 61901 and 64256 sub-shots in DEV and EVAL set, respectively.

The high-level feature detection task of TRECVID is to detect the presence or absence of 10 predetermined benchmark concepts in each shot of the EVAL set. The 10 semantic

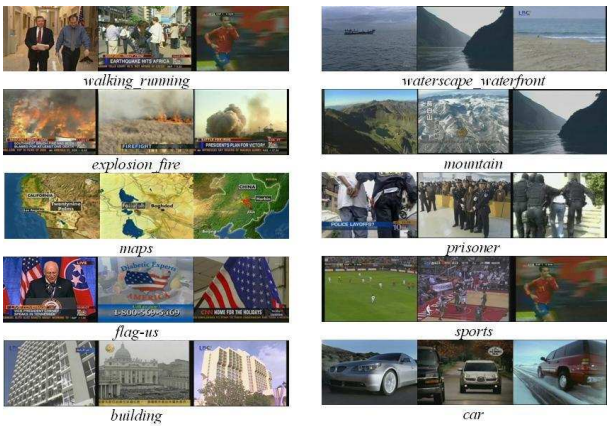


**Fig. 1**. The exemplary key-frames of the ten concepts.

concepts are: *walking_running, explosion_fire, maps, flag-US, building, waterscape_waterfront, mountain, prisoner, sports* and *car*. For each concept, systems are required to return ranked-lists of up to 2000 shots, and system performance is measured via non-interpolated mean average precision (MAP), which is a standard metric for document retrieval.

The low level features we used here are: 225-D block-wise color moments in LAB color space, which are extracted over $5\times5$ fixed grid partitions, each block is described using 9 dimensional features; 144-D color correlogram in HSV color space; 64-D color histogram in LAB color space. To avoid the *curse of dimensionality*, we separate the features into two set: Set 1 is with the color moments; Set 2 is with the correlogram and the histogram. Experiments are conducted on the two sets respectively and the results are finally combined through the linear fusion scheme in [7].

Using TCGRF method, the 64256 sub-shots in EVAL set are labeled as $f(subshot_i)$, and the sub-shots in the same shot are merged using the "max" rule:

$$f(shot_m) = max_{subshot_i \in shot_m}\{f(subshot_i)\} \quad (15)$$

Then the shot list is ranked according to $f(shot_m)$. We compare the experimental results of TCGRF and GRF [14] over the two feature sets respectively and the fusion results, which are shown in Table $1 \sim 3$. The evaluations are performed when all the parameters (for both TCGRF and GRF) are tuned to be nearly optimal by 5-fold cross-validations. From these comparisons, we can see that TCGRF improves GRF for 9 concepts out of 10 concepts, except for the concept *maps*. The improvement for the concepts *sports, walking_running* and *flag-US* is rather significant, while the improvement for *building* and *explosion_fire* is trivial. This is due to the fact that the temporal consistency in *sports, walking_running* and *flag-US* is much stronger than that in *explosion-fire, maps* and *building*.

**Table 1**. Comparisons of results over feature set 1

| Concept | GRF | TCGRF | Improvement |
|---|---|---|---|
| walking_running | 0.121 | 0.128 | + 20.7% |
| explosion_fire | 0.048 | 0.0484 | + 0.83% |
| maps | 0.458 | 0.458 | 0% |
| flag-US | 0.112 | 0.143 | + 27.7% |
| building | 0.4049 | 0.4051 | + 0.05% |
| waterscape_waterfront | 0.323 | 0.336 | + 4.02% |
| mountain | 0.304 | 0.324 | + 6.58% |
| prisoner | 0.00018 | 0.00047 | + 161% |
| sports | 0.265 | 0.369 | + 39.2% |
| car | 0.244 | 0.256 | + 4.92% |
| **MAP** | 0.228 | 0.249 | + 9.21% |

**Table 2**. Comparisons of results over feature set 2

| Concept | GRF | TCGRF | Improvement |
|---|---|---|---|
| walking_running | 0.099 | 0.126 | + 27.3% |
| explosion_fire | 0.035 | 0.035 | 0% |
| maps | 0.432 | 0.432 | 0% |
| flag-US | 0.032 | 0.039 | + 21.9% |
| building | 0.339 | 0.348 | + 2.65% |
| waterscape_waterfront | 0.306 | 0.329 | + 7.52% |
| mountain | 0.284 | 0.311 | + 9.51% |
| prisoner | 0.00023 | 0.0005 | + 117% |
| sports | 0.214 | 0.297 | + 38.8% |
| car | 0.166 | 0.2 | + 20.5% |
| **MAP** | 0.191 | 0.212 | + 11% |

**Table 3**. Comparisons of the fusion results

| Concept | GRF | TCGRF | Improvement |
|---|---|---|---|
| walking_running | 0.143 | 0.169 | + 18.2% |
| explosion_fire | 0.056 | 0.057 | + 1.79% |
| maps | 0.483 | 0.483 | 0% |
| flag-US | 0.129 | 0.156 | + 20.9% |
| building | 0.457 | 0.464 | + 1.53% |
| waterscape_waterfront | 0.351 | 0.364 | + 3.70% |
| mountain | 0.336 | 0.354 | + 5.36% |
| prisoner | 0.0003 | 0.0005 | + 66.7% |
| sports | 0.326 | 0.409 | + 25.5% |
| car | 0.257 | 0.266 | + 3.50% |
| **MAP** | 0.254 | 0.272 | + 7.09% |

## 5. CONCLUSIONS

A novel graph based method named TCGRF has been presented for automatic video semantic annotation. This method takes the advantage of the temporal consistency property of video data into graph based semi-supervised learning to improve the video annotation results. Experiments conducted on the TRECVID data set demonstrate that combining the temporal consistency into the graph based semi-supervised methods significantly improves the annotation performance compared with the normal graph based methods.

## 6. REFERENCES


[1] Guidelines for the TRECVID 2005 Evaluation. http://www-nlpir.nist.gov/projects/tv2005/tv2005.html.

[2] TREC-10 appendix on common evaluation measures. http://trec.nist.gov/pubs/trec10/appendices/measures.pdf.

[3] R. Angelova, G. Weikum. "Graph-based text classification: learn from your neighbors", *Proc. ACM Conference on Research & Development on Information Retrieval*, 2006.

[4] M. Belkin, P. Niyogi, V. Sindhwani. "Manifold Regularization: a Geometric Framework for Learning from Examples", *Journal of Machine Learning Research*, 2006.

[5] O. Chapelle, A. Zien, B. Scholkopf. *Semi-supervised Learning*, MIT Press, 2006.

[6] F. R. K. Chung. *Spectral Graph Theory*, American Mathematical Society, 1997.

[7] H. Tong, J. He, M. Li, C. Zhang, W.-Y. Ma. "Graph Based Multi-Modality Learning", *Proc. ACM Multimedia*, 2005.

[8] C. Wang, F. Jing, L. Zhang, H.-J. Zhang. "Image Annotation Refinement Using Random Walk with Restarts", *Proc. ACM Multimedia*, 2006.

[9] J. Tang, X.-S. Hua, G.-J. Qi, T. Mei, X. Wu. "Anisotropic Manifold Ranking for Video Annotation", *Proc. IEEE International Conference on Multimedia & Expo*, 2007.

[10] J. Yang, A. G. Hauptmann. "Exploring Temporal Consistency for Video Analysis and Retrieval", *ACM International Workshop on Multimedia Information Retrieval*, 2006.

[11] X. Yuan, X.-S. Hua, M. Wang, X. Wu. "Manifold-Ranking Based Video Concept Detection on Large Database and Feature Pool", *Proc. ACM Multimedia*, 2006.

[12] D. Zhou, O. Bousquet, et al. "Learning with Local and Global Consistency", *Proc. Neural Information Processing Systems*, 2003.

[13] X. Zhu. *Semi-Supervised Learning with Graphs*,PhD Thesis, CMU-LTI-05-192, May 2005.

[14] X. Zhu, Z. Ghahramani, J. Lafferty. "Semi-Supervised Learning Using Gaussian Fields and Harmonic Function", *Proc. International Conference on Machine Learning*, 2003.