DETECTING SEMANTIC CONCEPTS IN CONSUMER VIDEOS USING AUDIO

Junwei Liang[1], Qin Jin[1,2], Xixi He[1], Gang Yang[1], Jieping Xu[1], Xirong Li[1,2,3]*

1. Multimedia Computing Lab, School of Information, Renmin University of China, Beijing 100872
2. Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, 100872
3. Shanghai Key Laboratory of Intelligent Information Processing, 200443 China
{leongchunwai, qjin, xxlanmi, yanggang, xjieping，xirong}@ruc.edu.cn

ABSTRACT

With the increasing use of audio sensors in user generated content collection, how to detect semantic concepts using audio streams has become an important research problem. In this paper, we present a semantic concept annotation system using sound-tracks/audio of the video. We investigate three different acoustic feature representations for audio semantic concept annotation and explore fusion of audio annotation with visual annotation systems. We test our system on the data collection from HUAWEI Accurate and Fast Mobile Video Annotation Grand Challenge 2014. The experimental results show that our audio-only concept annotation system can detect semantic concepts significantly better than random guess. It can also provide significant complementary information to the visual-based concept annotation system for performance boost. Further detailed analysis shows that for interpreting a semantic concept both visually and acoustically, it is better to train concept models for the visual system and audio system using visual-driven and audio-driven ground truth separately.

Index Terms—Semantic Concept Annotation, Video Content Analysis, Audio Concept Analysis

1. INTRODUCTION

Current boom of user-generated content (UGC) on the Internet has attracted tremendous research interest in developing automatic technologies for organizing and indexing multimedia content [1]. The TRECVID annual evaluation organized by NIST has been an important benchmark [2]. With the increasing use of audio sensors in UGC data, semantic concept annotation using audio streams has become an important research problem. The audio information within the video can be very useful to detect semantic concepts, especially when the objects are hidden behind the camera and not appear in the visual content.

HUAWEI organized a grand challenge in the International Conference on Multimedia & Expo (ICME) 2014: HUAWEI Accurate and Fast Mobile Video Annotation Challenge [3]. The goal of this task is to analyze UGC videos and annotate their contents automatically. The labels to be annotated are 10 semantic concept classes, covering objects (e.g. “car”, “dog”, “flower”, “food” and “kids”), scenes (e.g. “beach”, “city view” and “Chinese antique building”) and events (“football” and “party”). The semantic concept annotation within the HUAWEI challenge is required to be at the frame-level. That means for each frame, we need to make a binary decision about the presence of a specific concept in the frame. Comparing to the semantic concept annotation task at the video level or supra segmental level in previous research, this task requires annotation with finer resolution and is a more challenging task. In this paper, we focus on detecting semantic concepts within UGC videos at frame level using audio information. We also investigate fusion of audio and visual annotation systems for additional performance improvement. Last but not least, we conduct further detailed analysis about how to best detect a semantic concept acoustically and visually.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 introduces the audio concept annotation system. Section 4 presents baseline experimental results. Section 5 presents further analysis and experimental results. Section 6 concludes the paper and describes potential future work.

2. RELATED WORK

The most related works are in soundtrack analysis and audio event classification. We summarize the previous research work from the following three focuses: (1) Number of sound classes. Much early works focused on detecting or distinguishing between a small number of sound classes such as speech, music, silence, noise, or applause. This was solved using various traditional machine learning and signal processing approaches [4-7]. (2) Quality of the audio data. Early work on audio event classification was largely done on sound databases [4] and clean broadcast or television program audio data [5]. Typical high quality database or broadcast data can be extremely clean, and “foreground” sounds are generally easy to distinguish from “background” sounds. The growing popularity of video sharing services such as YouTube, Dailymotion, Youku and Tudou in China etc. enables the vast increasing of user-generated videos. Analyzing such consumer videos is more challenging. (3): Granularity of the audio processing. We can roughly categorize the soundtrack analysis work into two categories: sub-soundtrack classification or entire soundtrack classification. Distinguishing between a small numbers of sound classes can be considered as a sub-soundtrack classification problem. It produces annotations of input data according to a fixed number of classes for which one has trained models. There also have been efforts to classify short audio clips with respect to the environment in which they were recorded [8]. The multimedia event detection (MED) using soundtrack is the entire soundtrack classification problem [9]. Modeling the event based on sub-soundtrack classification results has been one type of approaches in such tasks [9, 10]. Though the semantic indexing (SIN) task in TRECVID [2] has a subtask of localizing concepts on frame-level since 2013, we have not noticed any work that have used auditory method to help achieve the goal. Similar to the SIN subtask, the HUAWEI grand challenge can be categorized as a sub-soundtrack classification problem.

*Corresponding Author

3. AUDIO ANNOTATION SYSTEM DESCRIPTION

Our semantic concept annotation system using audio information only contains the following key components as shown in Figure 1: audio data pre-processing, audio feature extraction, concept annotation models and post-processing.

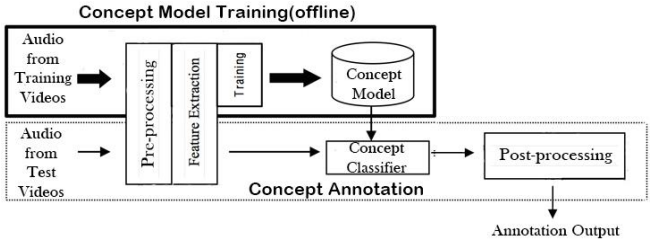

Figure 1. System Components

Pre-processing. In order to detect concept on frame-level, we chunk the audio stream into small segments with overlap (exp. 3-sec window and 1-sec shift), extract audio features and apply concept detection on those segments.

Audio Feature Representations. We explore different audio feature representations for concept annotation in this paper.

Bag-of-Words (BoW) features: The codebook model is a common technique used in the document classification (bag-of-words) [11] and image classification (bag-of-visual words) [12]. The similar bag-of-audio-words model has also been applied in the sound track analysis work [13-14]. In our system, we use bag-of-audio-words model to represent each audio segment by assigning low-level acoustic features to a discrete set of codewords in the vocabulary (codebook) thus providing a histogram of codewords' counts. These codewords are learnt via unsupervised clustering. The discriminative power of such a codebook is governed by the size of the codebook and by the assignment of features to codewords [7]. In this paper we apply this model to the low level MFCC features. The MFCC features are computed every 25ms with 10ms shift and are in 39 dimensions (13MFCC + 13delta + 13ddelta). The codebook is learnt by applying kmeans clustering algorithm with K=4096 on the whole training dataset. Each audio segment is then represented as a distribution over these 4096 codewords' by using soft-assignment (by adding the closest 5 codewords' count) of MFCC features to these codewords'.

BoW+TF-IDF features: In the previous feature representation, when we calculate the bag-of-audio-word features, we only consider the hard counts of each codeword (as term frequency). Some codeword may be common noise that may not be useful in classification. Therefore we consider using the term frequency–inverse document frequency (tf-idf) method to eliminate the influence of such noises, similar to the work in [15-16]. For each codeword, we calculate its inverse document frequency in the training set and then multiply it with the original term frequency in all the dataset and get the IDF-bag-of-audio-word features.

Gaussian Super Vector Features: Instead of the bag-of-audio-word representation, we experiment another method to represent low-level MFCC features. Gaussian supervectors (GSV) have been successfully used on the speaker verification task [17]. A GSV is constructed by stacking the means, diagonal covariances, and/or component weights of the mixture model. We first trained a universal background model (UBM) by sampling audio from the training set. To generate the GSV feature representation for each audio segment, we first MAP adapt to the UBM based on the MFCC features extracted from this segment and then create a super vector by concatenating the means of each Gaussian component in the adapted GMM.

Concept Annotation Models. After we extract the audio features, we train two-class SVM classifiers for each of the 10 concepts. As the training data is overwhelmed by negative examples (Table 1), we train classifiers with the Negative Bootstrap algorithm [18]. The algorithm takes a fixed number (N) of positive examples and iteratively selects negative examples which are most misclassified by current classifiers (N=3000 in this paper). The algorithm randomly samples 10xN number of negative examples from the remaining negative examples as candidates at each iteration. An ensemble of classifiers trained in the previous iterations is used to classify each negative candidate examples. The top N most misclassified candidates are selected and used together with the N positive examples to train a new classifier. In order to improve the efficiency of the training process, we use Fast intersection kernel SVMs (FikSVM) as reported in [19].

Post-processing. Intuitively, if a concept occurs within a video, it is usually not an instantaneous appearance. It normally lasts for certain duration. Therefore, we conduct boundary padding and cross-segment smoothing over the raw annotation results. We expand the beginning and ending of the detected segments. We also merge two detected segments if they belong to the same concept and the gap between them is below a certain threshold (padding=3sec, gap threshold=3sec in this paper).

4. BASELINE EXPERIMENTS

4.1 Database Description

The HUAWEI dataset contains 2,666 UGC videos with frame-level ground truth provided for ten concepts. The video resolutions and frame per second (fps) vary among all videos. We divide the dataset into a training set (1300 videos), a development set for tuning model parameters and fusions weights (477 videos) and a test set (886 videos). The ground truth label files provide the exact frame index of each concept that appears within videos. Table 1 shows the total number frames of positive and negative examples for each concept. The amount of negative examples is overwhelmingly larger than the amount of positive examples.

Table 1: number of frames for positive and negative examples

Concepts	#pos frms	#neg frms	%pos
beach	664793	9340147	6.6%
car	1157352	8847588	11.6%
ch_bldg	772805	9232135	7.7%
city view	801583	9203357	8.0%
dog	520744	9484196	5.2%
flower	1082986	8921954	10.8%
food	378046	9626894	3.8%
fb-game	1604012	8400928	16.0%
kids	1525771	8479169	15.3%
party	780240	9224700	7.8%

4.2 Baseline Experimental Results

We use the average precision to evaluate the concept annotation performance for each concept class:

$$AP = \frac{1}{R}\sum_{j=1}^{n} I_j \times \frac{R_j}{j} \quad (1)$$

where R is total number of relevant segments of that concept, n is the total amount of segments, $I_j=1$ when the j^{th} segment is relevant

otherwise $I_j=0$. R_j is the number of relevant segments in the first j segments.

Table 2 shows the baseline results with the three different feature representations. The Audio annotation system with BoW features yields 29.8% of mean AP over all 10 concepts, 28.8% with BoW+TF-IDF features, and 29.7% with GSV features. From the results, we can see that the concept annotation based on audio only achieves significantly better performance than random guess. The system with BoW+TF-IDF features does not get any gain over BoW features, which indicates that adding inverse-document-frequency at the feature level does not help. We suspect that the SVM classifier may already compensate the inverse-document-frequency implicitly. We then fuse the three audio annotation baseline systems via late fusion. We use a coordinate ascent algorithm [20] to find the optimal fusion weights on the development data and apply them on the test data. The fusion of three baseline systems achieves additional improvement (boost mean AP to 33.6%), which shows the three different audio features are complementary at certain level. We also conduct pair-wise fusion between any two of the three baseline systems; their performance is comparable but slightly worse. From the results, we can also see that some concept classes, which are acoustically easy to distinguish such as "football game", "dog", "kids", "party" clearly achieve much better performance than others (with AP of 75.3%, 47.8%, 47.1%, and 36.9% respectively).

Since significant semantic information is conveyed in the visual stream, we also develop the concept annotation system using visual information (SVM classifier of same structure as in audio system trained with 1x1+1x3 SIFT BoW features). Intuitively, the audio and visual streams contain complementary information for interpreting a semantic concept. We therefore explore to combine the fused audio system and the visual system. As shown in Table 2, although the visual system achieves much better performance than the audio system, combining them achieves improvement over all of the 10 concept classes. On average the fusion of audio and visual system achieves a relative improvement of 13.6% over mean AP (boost mean AP from 63.2% to 68.2%). The fusion weight assigned to the audio system is listed in the last column in Table 2.

Table 2: Baseline concept annotation system performance

Concept	BoW Audio Sys	Tf-idf Audio Sys	GSV Audio Sys	Audio Fusion	Visual Sys	Audio & Visual fusion	Audio fusion weight
beach	12.2%	11.8%	14.8%	***15.2%***	60.3%	**61.9%**	0.2
car	25.9%	26.0%	26.0%	***28.3%***	65.5%	**66.1%**	0.1
ch-bldg	18.4%	17.1%	16.7%	***22.2%***	65.1%	**68.6%**	0.3
cityview	21.5%	20.2%	18.8%	***23.2%***	57.2%	**60.5%**	0.3
dog	46.0%	44.5%	46.4%	***47.8%***	49.7%	**66.3%**	0.5
flower	27.8%	27.2%	26.2%	***31.8%***	74.6%	**76.9%**	0.2
food	7.3%	7.3%	7.6%	***8.6%***	46.4%	**46.7%**	0.3
fb-game	71.5%	71.3%	69.4%	***75.3%***	97.3%	**97.9%**	0.3
kids	41.7%	39.4%	37.4%	***47.1%***	38.3%	**56.6%**	0.6
party	25.2%	23.2%	33.5%	***36.9%***	77.5%	**80.9%**	0.6
Average	**29.8%**	**28.8%**	**29.7%**	***33.6%***	**63.2%**	**68.2%**	-

5. FURTHER DETAILED ANALYSIS

We inspect baseline results in detail and notice that in some test videos the system detects certain acoustically salient concepts but they are not labeled in the ground truth. For example, for a segment in the video tr0209.mp4 with kids talking behind the camera about the chicken pacing in the front, audio-only system successfully detects "kids", while the ground truth file does not label such concept. This indicates that the ground truth is generated mainly based on the visual evidences. Therefore much audio semantic content is left out in the HUAWEI UGC data collection. However, for a semantic concept that can relate to both visual and acoustic evidences, when the visual system fails to detect, audio annotation system will be the best solution. We therefore come up with the following additional experiments and analysis.

5.1. Audio-driven Concept Ground Truth

As we look more closely into the ground truth files, we believe that the provided manual labels are based on visual content. For example, the videos with dog's image in the foreground and kid's talking behind the camera are only labeled with "dog" but no "kids". In order to further investigate semantic concept annotation from both audio and visual point of view, we need consistent ground truth with both audio and visual evidences. We therefore choose 6 acoustically salient concepts (kids, football-game, dog, party, car, beach), and hand label the whole dataset by listening to the sound tracks without looking at the videos to generate the new audio-driven semantic concept ground truth. We compare the original visual-driven ground truth and the new audio-driven ground truth in Table 3. The "Visual" and "Audio" columns show the number of frames labeled as each concept in the original visual-driven ground truth and the new audio-driven ground truth respectively. The "Intersect" column shows how many frames contain both visual and audio content for each semantic concept (intersection of visual-driven and audio-driven ground truth). The "%Visual" and "%Audio" columns show the percentage of frames in visual-driven ground truth and audio-driven ground truth that actually contain both visual and audio contents for the semantic concept respectively. For example, for the concept "kids", 91% of the frames labeled with "kids" in the audio-driven ground truth are associated with visual labels, while only 43% of frames labeled as "kids" in the visual-driven ground truth are related to "kids" acoustically. This indicates that there are many videos with kids appearing in the images but without kids' voice; however if the videos do contain the kids' voice they most likely contain kids' images at the same time. From the statistics we may infer that the audio evidences for a concept are more likely associated with visual evidences.

Table 3. Comparison of original visual-driven ground truth and new audio-driven ground truth

Concept	Visual	Audio	Intersect	%Visual	%Audio
kids	1525771	726902	662745	43%	91%
party	780240	972960	731453	94%	75%
car	1157352	1151423	533031	46%	46%
fb-game	1604012	1117904	1114555	69%	99%
beach	664793	359424	289476	44%	81%
dog	520744	123694	121888	23%	99%

We train and test our audio system based on the new audio-driven ground truth as well. We only focus on the 6 acoustically more salient concepts. For all experiments and analysis in the following sections, we only use the BoW Audio baseline system. Table 4 compares the performance of the original BoW baseline system, the BoW baseline system scored against the new audio-driven ground truth, and the re-trained BoW system based on the new audio-driven ground truth plus tested on the new ground truth.

The results show that re-scoring the original BoW baseline system on the new audio-driven ground truth gets better performance for two of the concepts ("kids", "party"). This agrees with our intuition because "kids" and "party" are more acoustically consistent even when the ground truth is generated visually. However the audio content for other concepts within the visual-driven ground truth is very much inconsistent. Re-train the system with the new audio-driven ground truth further improves the performance by increasing the mean AP from 35.3% to 43.1%.

Table 4. Evaluating audio BoW baseline system on the newly constructed audio-driven ground truth

Concept	Original Baseline	Baseline Re-scored	Re-trained + Re-scored
kids	41.7%	***44.8%***	52.8%
party	25.2%	***34.1%***	35.0%
car	25.9%	24.8%	43.6%
fb-game	71.5%	62.9%	75.2%
beach	12.2%	9.9%	12.5%
dog	46.0%	35.1%	43.6%
Average	**37.8%**	**35.3%**	**43.1%**

5.2. Pure Music Videos

We also notice that quite some videos in the test set are edited with pure music (such as a car video with pure pop music), which isn't really acoustically related to the visual semantic concept; but our audio system may classify them as the candidate concepts, which will increase errors. If we filter out such pure music videos from the test set, we can get relative improvement of 11.6% (increasing average mean AP from 43.1% to 49.7%). Therefore in the future, we will consider building a music classifier to detect incoming pure music videos automatically.

5.3. Fusion of Audio and Visual Systems

As shown in previous baseline results, combining the visual annotation system with audio annotation system achieves nice improvement over all the 10 classes. To further investigate the complementary information between the audio content and the visual content, we conduct more fusion experiments of these two systems. In the following experiments, we want to evaluate the fusion performance on the contents that are both visually and acoustically identifiable, we therefore use the intersection of audio-driven and visual-driven ground truth (as shown in Table 3) as scoring reference. Table 5 shows the audio system and visual system performance scored against the intersection ground truth. "Audio 1" refers to the BoW audio baseline system trained using the visual-driven ground truth. "Audio 2" refers to the BoW audio system trained using the audio-driven ground truth. "Visual 1" refers to the baseline visual system trained using the visual-driven ground truth. We also train a new visual and a new audio system using the intersection ground truth. "Visual 2" and "Audio 3" refer to these two new systems respectively. "Visual 1" and "Audio 2" are the best performing visual and audio systems respectively, which infers that it is best to train visual and audio systems using ground truth from their own perspective. We refer to these systems using abbreviations A1, A2, A3, V1, and V2 to save space.

We conduct four fusion experiments: fusion of "A1" and "V1" (named Fusion I), "A2" and "V1" (named Fusion II), "A2" and "V2" (named Fusion III), and "A3" and "V2" (named Fusion IV). Detailed fusion results are shown in Table 6. All fusions improve the corresponding single visual-only or audio-only system, which again proves that audio and visual streams contain complementary information for interpreting a semantic concept. A higher relative improvement of 20.2% from Fusion I (compared to 13.6% from baseline fusion in Table 2) indicates that for both visually and acoustically relevant content, fusion brings more gain. Fusion II achieves better improvement than Fusion I, which indicates that it is better to train audio semantic concept using audio-driven ground truth. Fusion III & IV cannot out-perform Fusion II shows that training visual semantic concept is better to use visual-driven ground truth. Fusion II achieves the best relative improvement of 33.5%, in which visual and audio systems are trained based on visual-driven and audio-driven ground truth independently.

Table 5. Performance scored against intersection ground truth

Concept	Audio 1	Audio 2	Audio 3	Visual 1	Visual 2
kids	47.4%	54.1%	51.5%	26.5%	23.8%
party	34.5%	35.7%	34.2%	80.8%	80.2%
car	17.6%	20.6%	19.4%	37.4%	39.5%
fb-game	77.4%	80.2%	80.4%	94.5%	94.5%
beach	11.4%	13.0%	12.5%	44.7%	42.0%
dog	40.9%	49.9%	50.4%	13.3%	12.8%
Average	**38.2%**	**42.2%**	**41.4%**	**49.5%**	**48.8%**

Table 6. Fusion of Audio and Visual concept annotation systems

Concept	Fusion I (A1+V1)	Fusion II (A2+V1)	Fusion III (A2+V2)	Fusion IV (A3+V2)
kids	47.9%	61.1%	59.6%	57.5%
party	82.9%	83.4%	82.9%	82.4%
car	38.3%	45.9%	46.6%	45.8%
fb-game	95.7%	97.0%	96.4%	96.3%
beach	50.9%	55.9%	52.1%	51.7%
dog	42.7%	55.3%	51.6%	51.8%
Average	**59.7%**	**66.4%**	**64.9%**	**64.3%**

6. CONCLUSIONS

This paper presents our semantic concept annotation system using audio only. The system uses three different audio feature representations and negative bootstrap SVM concept classifier. The experimental results on the HUAWEI grand challenge UGC video data show that our audio-only concept annotation system can detect semantic concepts significantly better than random guess. When combining with visual-only concept annotation system, it brings improvement in general over all concepts and more significantly on certain concepts. Further detailed analysis shows that it is better to interpret a concept both visually and acoustically via training visual system and audio system using visual-driven and audio-driven labels separately. A relative improvement of 33.5% is achieved when fusing the audio and visual systems according to such criteria. In the future work, we will explore automatic approaches for detecting music edited videos and investigate the potential of utilizing the concept co-occurrence property.

ACKNOWLEDGEMENTS

This work is supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01), the Beijing Natural Science Foundation (No. 4142029), NSFC (No. 61303184), SRFDP (No. 20130004120006), and Shanghai Key Laboratory of Intelligent Information Processing, China (Grant No. IIPL-2014-002).

6. REFERENCES

[1] Snoek, C. and Worring, M.: Concept-based Video Retrieval. Foundations and Trends in Information Retrieval, 2009.

[2] Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Kraaij, W., Smeaton, A.F., Quéenot, G.: TRECVID 2013 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In: Proceedings of TRECVID 2013, NIST, USA. http://wwwnlpir.nist.gov/projects/tvpubs/tv13.papers/tv13overview.pdf.

[3] ICME 2014 Huawei Accurate and Fast Mobile Video Annotation Challenge http://www.icme2014.org/huawei-accurate-and-fast-mobile-video-annotation-challenge.

[4] Wold, E., Blum, T., Keislar, D., and Wheaten, J.: Content-based Classification, Search, and Retrieval of Audio. In: IEEE Multimedia, 3(3), 1996.

[5] Saunders, J.: Real-time Discrimination of Broadcast Speech/Music. In: ICASSP, 1996.

[6] Scheirer, E. and Slaney, M.: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In: ICASSP, 1997.

[7] Williams, G. and Ellis, D.P.W.: Speech/Music Discrimination Based on Posterior Probability Features. In: Eurospeech, 1999.

[8] Ma, L., Milner, B., and Smith, D.: Acoustic Environment Classification. In: ACM Transactions on Speech and Language Processing, 3(2), 2006.

[9] Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., and Huopaniemi, J.: Audio-based Context Recognition. In: IEEE Trans. on Audio, Speech, and Language Processing, 14(1), 2006.

[10] Brown, L. et al.: IBM Research and Columbia University TRECVID-2013 Multimedia Event Detection (MED), Multimedia Event Recounting (MER), Surveillance Event Detection (SED), and Semantic Indexing (SIN) Systems. In: TRECVID Workshop, 2013.

[11] Xue, X.B.; Zhou, Z.H.: Distributional Features for Text Categorization. In: IEEE Transactions on Knowledge and Data Engineering, 21(3), 2008.

[12] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR 2007.

[13] Lee, K. and Ellis, D.P.W.: Audio-Based Semantic Concept Classification for Consumer Video. In: IEEE Transactions On Audio, Speech, and Language Processing, 18(6), 2010.

[14] Jin., Q., Schulam, F., Rawat, S., Burger, S., Ding, D，Metze, F.: Categorizing Consumer Videos Using Audio. In: Interspeech, 2012.

[15] Pancoast, S., Akbacak, M. “N-gram extension for bag-of-audio-words”, Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on DOI: 10.1109/ICASSP.2013.6637754 Publication Year: 2013, Page(s): 778 - 782

[16] Pancoast, S., Akbacak, M. “Softening quantization in bag-of-audio-words”, Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on DOI: 10.1109/ICASSP.2014.6853821 Publication Year: 2014 , Page(s): 1370 - 1374.

[17] Campbell, W.M., Sturim, D.E. and Reynolds, D.A. “Support vector machines using GMM supervectors for speaker verification”, IEEE Signal Processing Letters, 2006, pp 308-311.

[18] Li, X., Snoek, C., Worring, M., Koelma, D., Smeulders, A.: Bootstrapping Visual Categorization With Relevant Negatives. In: IEEE Transactions on Multimedia, 15(4), 2013.

[19] Maji, S., Berg, A., Malik, J.: Classification using intersectional kernel support vector machines is efficient. In: CVPR 2008.

[20] Xirong Li, Cees Snoek, Marcel Worring, Arnold Smeulders, Fusing concept detection and geo context for visual search, ICMR 2012.