THE ROLE OF VISUAL CONTENT AND STYLE FOR CONCERT VIDEO INDEXING

C.G.M. Snoek, M. Worring, A.W.M. Smeulders

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ
Amsterdam, The Netherlands
{cgmsnoek,worring,smeulders}@science.uva.nl
http://www.mediamill.nl

B. Freiburg

Stichting Fabchannel
Weteringsschans 6–8, 1017 SG
Amsterdam, The Netherlands
bauke@fabchannel.com
http://www.fabchannel.com

ABSTRACT

This paper contributes to the automatic indexing of concert video. In contrast to traditional methods, which rely primarily on audio information for summarization applications, we explore how a visual-only concept detection approach could be employed. We investigate how our recent method for news video indexing – which takes into account the role of content and style – generalizes to the concert domain. We analyze concert video on three levels of visual abstraction, namely: content, style, and their fusion. Experiments with 12 concept detectors, on 45 hours of visually challenging concert video, show that the automatically learned best approach is concept-dependent. Moreover, these results suggest that the visual modality provides ample opportunity for more effective indexing and retrieval of concert video when used in addition to the auditory modality.

1. INTRODUCTION

People enjoy music traditionally by listening. Sounds reach the ear by means of live acts, personal audio devices, and the Internet. Due to improved hardware capabilities and ever increasing broadband connections, music is often accompanied by carefully produced visual information in the form of concert footage and clips. Besides listening, people enjoy music nowadays by watching also. Since the size of multimedia music collections is on the rise, there is a clear need for automatic indexing and search tools. Most research in music retrieval emphasizes an audio-only approach; see e.g. [2, 12] for a collection of state-of-the-art developments. Surprisingly, only few works in literature consider the fact that music often has a visual component; representative exceptions are [1,5,8]. These works exploit the visual channel as a secondary aid to fine tune audio-based segmentation and summarization. We question why the visual stream is not used as a complementary modality. In this paper we, therefore, explore the utility of the visual modality for the semantic indexing of concert video.

Semantic video indexing has been explored on domains like sports and news; especially as part of the TRECVID benchmark [9]. Systems index news video at the granularity of a shot, i.e. a continuous spatiotemporal camera action, with concepts like anchor, outdoor, and airplane [6, 10]. The question arises whether those existing techniques for concept detection generalize to music video. To arrive at generic video indexing, we departed in [10] from the premise that the essence of produced video, like a concert video or broadcast news, is that an author creates the final program [3]. It is more than just the content. Before creation, the author starts with a semantic idea: an interplay of concepts like people, objects, settings, and events. To stress the semantics of the message, guiding the audience in its interpretation, the author combines various stylish production facets, such as camera framing. Hence, the core of semantic video indexing is to inverse this authoring process. We showed in [10] that generic indexing of concepts in news video is feasible indeed when analysis adheres to this authoring metaphor, i.e. exploiting the fact that news video is authored by taking the role of content and style into account.

In this paper, we investigate whether the authoring metaphor generalizes to the domain of music video. We focus specifically on concert video registrations as these music videos have a high consistency in production style while simultaneously posing severe challenges for visual content analysis. These challenges are caused by the fact that footage is typically recorded in relatively dark settings with large amounts of camera motion and various light effects. Given these challenges, applying the authoring metaphor to concert video is a non-trivial extension. Hence, we need to reconsider the role of visual content and visual style. We develop 12 concept detectors for concert video and we empirically investigate the role of visual content, style, and their fusion.

2. CONCEPT DETECTORS FOR CONCERT VIDEO

In this section we detail concert concepts, and how to detect them automatically using analysis of visual content, visual

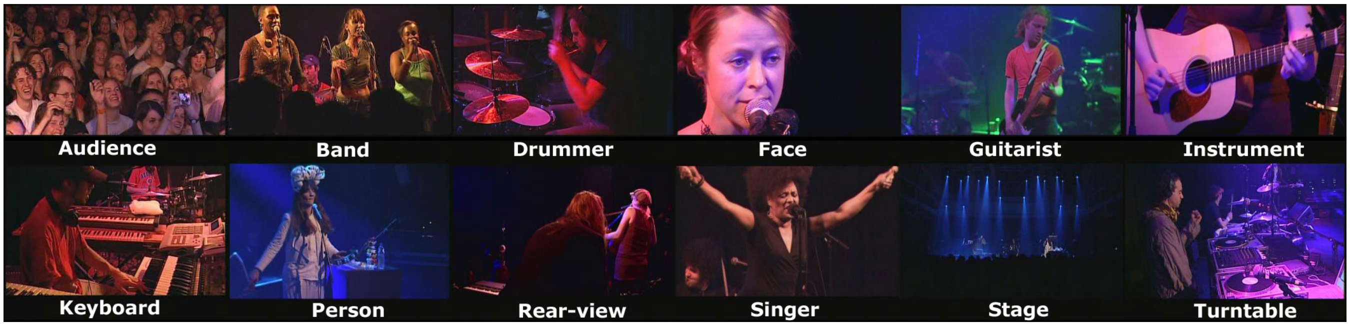

Fig. 1. Visual impression of 12 common concert concepts that we aim to detect in this paper using analysis of visual content, visual style, and their fusion. Note the challenging nature of the video data, since it is recorded in relatively dark settings with large amounts of camera motion and various light effects.

style, and their fusion.

2.1. Concert Concepts

In contrast to news video, where the number of concepts is unrestricted, the number of concepts that may appear in a concert is more or less fixed. A band plays on stage for an audience. Thus, major concepts are related to the role of the band members, e.g. lead singer, or guitarist, and the type of instruments that they play, e.g. drums or keyboard. Although quite many instruments exist, most bands typically use guitars, drums, and keyboards. We chose 12 concert concepts based on an interview with concert producers, previous mentioning in literature [5], and expected utility for concert video users. These 12 concert concepts are depicted in Fig. 1.

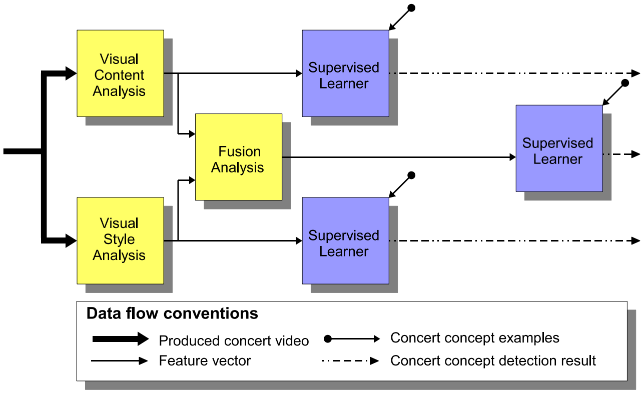

Fig. 2. Framework for semantic indexing of concert video using analysis of visual content, visual style, and their fusion.

2.2. Analyzing Concert Video

We employ the framework developed in [10] as guiding principle to arrive at concert concept detectors. Given a feature vector $\vec{x}_i$, part of a shot i, the aim is to obtain a confidence measure, $p(\omega_j|\vec{x}_i)$, which indicates whether concert concept ω_j is present in a shot. Feature extraction methods in the framework address visual content analysis, visual style analysis, and their fusion [10]. We rely on supervised machine learning to convert a feature vector to a confidence measure, based on concert concept examples. The framework is detailed in Fig. 2. We stress that our framework uses existing implementations tuned for news video. What differs is the used training set, test set, and the examples needed for learning detectors. Note that we have not performed any optimization to fine tune results for the domain of concert video. Therefore, we explain the implementation only briefly; where needed we provide pointers to published papers covering in-depth technical details.

Visual Content Analysis is based on the method described in [4]. In short, the procedure first extracts a number of color invariant texture features per pixel. Based on these features, it labels a set of predefined regions in a key frame image with similarity scores for a cluster of 15 low-level visual concepts. This yields a vector, where each element represents a similarity score to one of the 15 regional concept clusters. We vary the size of the predefined regions to obtain a total of 8 concept occurrence vectors that characterize both global and local color-texture information. We concatenate the vectors to yield a 120-dimensional visual content vector per key frame, $\vec{c}_i$. To learn concepts, $\vec{c}_i$ serves as the input for the supervised learner.

Visual Style Analysis uses a subset of the detectors proposed in [10]. Here we provide a summary of the visual detectors only. We compute the camera distance from the size of detected faces [7]. It is undefined when no face is detected. In addition to camera distance, several types of camera work are detected, e.g. pan, tilt, zoom, and so on. Finally, we also estimate the amount of camera motion. We have chosen to convert the output of all visual style detectors to an ordinal scale, as this allows for easy fusion into visual style vector $\vec{s}_i$. To learn semantic concepts, $\vec{s}_i$ serves as the input for the supervised learner.

Fusion Analysis combines the feature vectors resulting from content and style analysis. We adopt the fusion method proposed in [10], using vector concatenation to unite the features $\vec{c}_i$ and $\vec{s}_i$ into fusion vector $\vec{f}_i$. To learn semantic concepts, $\vec{f}_i$ serves as the input for the supervised learner.

Supervised Learner obtains confidence measure $p(\omega_j|\vec{x}_i)$. We choose the Support Vector Machine (SVM) framework, which has proven to be a solid choice [4, 6, 8, 10]. Here we use the LIBSVM implementation with radial basis function and probabilistic output. Classifiers thus trained for ω_j, result in an estimate $p(\omega_j|\vec{x}_i)$. We obtain good SVM parameter settings by performing an iterative search on a large number of combinations on training data. We select the parameters with the best performance after 3-fold cross validation, resulting in $p^*(\omega_j|\vec{x}_i)$. We apply the concept detectors on the test set and rank concept detection results based on $p^*(\cdot)$.

3. EXPERIMENTAL SETUP

3.1. Video Data

We use concert video registrations from Fabchannel to evaluate our approach. Fabchannel currently narrowcasts over 700 live concert music videos from the Paradiso and Melkweg club venues in Amsterdam over the Internet. For our experiments we selected a subset, consisting of 38 full-length video registrations, that covers a wide diversity in genre, i.e. Dance, Metal, Singer/Songwriter, HipHop, Rock, and Punk . The concert videos are from artists like *Spinvis*, *Aerogramme*, *Millencolin*, and *Daughters of Soul*. All concerts are recorded in MPEG1 between April 2005 and February 2006 with a total length of 45 hours. We use a standard shot segmentation tool to segment the videos. The training set contains 25 concerts (24,231 shots), the test set contains the remaining 13 concerts (16,880 shots). All videos are also viewable on Fabchannel.

3.2. Concept Detector Experiments

Since supervised learning of concept detectors requires labeled samples, we manually annotated the key frames in the training set for each of the 12 concert concepts defined in Section 2.1. Presence of a concept was assumed to be binary, i.e. it is either visible during a shot/key frame or not. We carry out three experiments. In experiment 1 we investigate the role of visual content on concert concept detection performance. This is followed by visual style in experiment 2. Finally, in experiment 3 we explore the role of fusion.

3.3. Evaluation

To determine the accuracy of concept detector rankings we use *precision at n*. This value gives the fraction of correctly annotated shots within the first n retrieved results. Note that this measure assumes there are more than n relevant shots per concept, for sparse concepts this is not necessarily the case.

3.4. Concert Video Search Demo

To demonstrate the potential of our approach, we developed a concert video search engine. It allows for query by concert concept, see Fig. 3, to let users search for footage of favorite band members for example. The system displays results in a cross browser [11], see Fig. 4.

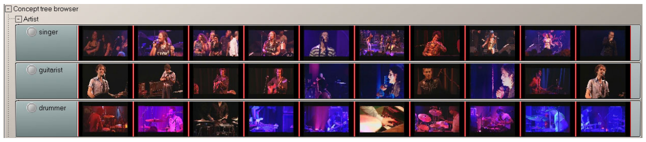

Fig. 3. Detail of the query panel of our concert video search engine, showing top 10 indexed results for three concert concepts.

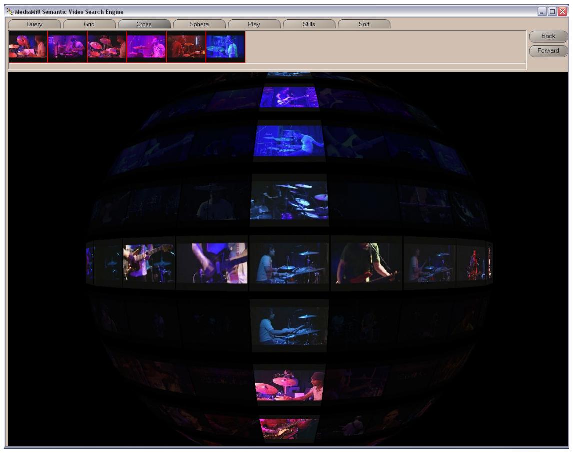

Fig. 4. Cross browser [11] showing from top to bottom ranked results for drummer, and from left to right the time line of the concert.

4. RESULTS

We compare the influence of visual content analysis, visual style analysis, and their fusion on concept detection performance. We present the results, with varying precision at n, in Table 1. Visual content analysis obtains the best performance overall. Compared to style and fusion, content analysis works particularly well for concepts emphasizing band members and their instruments, e.g. *keyboard*, *drummer*, and *guitarist*. Results for style analysis show that three concepts are detected with good performance: *singer*, *person*, and *face*. For these concepts the camera distance is a robust feature. Since these rely mainly on detected faces, style analysis does not perform well for concepts where faces are absent. The combination of style and content features shows the best result for only one concept: *stage*. For the other concepts, the combination

Table 1. Precisions at 10, 20, 50 and 100 per concert concept for visual content analysis, visual style analysis and their fusion.

Concert Concept	Exp. 1: Visual Content Analysis				Exp. 2: Visual Style Analysis				Exp. 3: Fusion Analysis			
	p@10	p@20	p@50	p@100	p@10	p@20	p@50	p@100	p@10	p@20	p@50	p@100
Audience	0.20	0.30	0.32	0.22	0.10	0.05	0.04	0.02	0.10	0.10	0.04	0.06
Band	0.90	0.85	0.68	0.67	0.70	0.65	0.56	0.49	0.80	0.75	0.68	0.66
Drummer	0.70	0.65	0.64	0.62	0.20	0.10	0.14	0.17	0.20	0.25	0.28	0.34
Face	1.00	0.95	0.92	0.93	1.00	0.95	0.98	0.97	1.00	1.00	1.00	0.95
Guitarist	0.50	0.45	0.50	0.35	0.10	0.15	0.14	0.19	0.00	0.10	0.18	0.22
Instrument	0.80	0.80	0.66	0.63	0.30	0.25	0.18	0.18	0.70	0.70	0.56	0.54
Keyboard	0.20	0.25	0.24	0.24	0.00	0.00	0.00	0.01	0.10	0.15	0.08	0.11
Person	0.80	0.75	0.68	0.69	0.90	0.85	0.82	0.81	0.60	0.60	0.64	0.67
Rear-view	0.60	0.50	0.52	0.44	0.00	0.00	0.12	0.11	0.20	0.20	0.16	0.22
Singer	0.70	0.65	0.58	0.53	0.80	0.65	0.70	0.71	0.40	0.40	0.48	0.61
Stage	0.70	0.80	0.76	0.71	0.60	0.50	0.52	0.53	0.90	0.75	0.76	0.80
Turntable	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean	*0.59*	*0.58*	*0.54*	*0.50*	*0.39*	*0.35*	*0.35*	*0.36*	*0.42*	*0.42*	*0.50*	*0.43*

seems to be the average of the two analysis approaches in isolation. Apparently the current implementation of content and style features do not complement each other. Because only a limited number of training samples were available for *turntable*, none of the analysis approaches works well for this concept. Taking all results into account, we observe a tendency in the precision at n results. The results of visual content analysis seem to decrease towards a higher depth n. Yet, the style and fused analysis seem to be stable, albeit lower on average. We are currently investigating at what precision the break even point resides.

5. CONCLUSION

In this paper, we explore the role of visual content, visual style, and their fusion for semantic indexing of concert video. Specifically we investigate whether our proposed framework for news video indexing generalizes to the visually challenging domain of concert video. Experiments with a lexicon of 12 semantic concepts on 45 hours of narrowcast concert video demonstrate that this is indeed the case. Visual content analysis performs better when the classification depends more on visual details like instruments. In contrast, visual style analysis should be used when the semantic concept is detectable based on such features as camera distance. Our results indicate no synergetic effects can be contributed to a combination of content and style. Naturally the results can be improved further by inclusion of the auditory modality and more advanced fusion schemes, which we aim to evaluate in future research.

Acknowledgement

This research is sponsored by the BSIK MultimediaN project. The authors thank UvA colleagues Joeri Edens (experiments), Jan van Gemert (features), and Ork de Rooij (interface).

6. REFERENCES

[1] L. Agnihotri, N. Dimitrova, and J. R. Kender. Design and evaluation of a music video summarization system. In *Proc. IEEE ICME*, pages 1943–1946, Taipei, Taiwan, 2004.

[2] X. Amatriain et al., editors. *Proceedings AMCMM Workshop*. Santa Barbara, USA, 2006.

[3] D. Bordwell and K. Thompson. *Film Art: An Introduction*. McGraw-Hill, NY, USA, 5th edition, 1997.

[4] J. van Gemert et al. Robust scene categorization by learning image statistics in context. In *SLAM Workshop, in conjunction with CVPR'06*, New York, USA, 2006.

[5] Y. van Houten et al. The MultimediaN concert video browser. In *Proc. IEEE ICME*, Amsterdam, The Netherlands, 2005.

[6] M. Naphade and J. Smith. On the detection of semantic concepts at TRECVID. In *ACM Multimedia*, NY, USA, 2004.

[7] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *IJCV*, 56(3):151–177, 2004.

[8] X. Shao et al. Automatic summarization of music videos. *ACM TOMCCAP*, 2(2):127–148, 2006.

[9] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. ACM MIR*, pages 321–330, 2006.

[10] C. Snoek et al. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE TPAMI*, 28(10):1678–1689, 2006.

[11] C. Snoek et al. A learned lexicon-driven paradigm for interactive video retrieval. *IEEE TMM*, 9(2):280–292, 2007.

[12] G. Tzanetakis et al., editors. *Proc. ISMIR*. Victoria, Canada, 2006.