Multi-modal Video Concept Detection for Scalable Logging of Pre-Production Broadcast Content

C. Gray, J. Collomosse
Centre for Vision Speech and Signal Processing
University of Surrey
Guildford, United Kingdom
{charles.gray | j.collomosse}@surrey.ac.uk

J. Thorpe and A. Turner
Broadcast & Professional Research Labs
Sony Professional Solutions Europe
Basingstoke, United Kingdom
{jonathan.thorpe | alan.turner}

ABSTRACT

We present a scalable semantic video concept detection framework, applied to automated metadata annotation (video logging) in a broadcast production environment. Video logging demands both accurate and fast concept detection. Whilst research often focuses on the former, the latter is essential in practical scenarios where days of footage may be shot per broadcast episode and production is dependent on immediate availability of metadata. We present a hierarchical classification framework that delivers benefits to both through two contributions. First, a dynamic weighting scheme for combining video features from multiple modalities enabling higher accuracy detection rates over diverse production footage. Second, a hierarchical classification strategy that exploits ontological relationships between concepts to scale sub-linearly with the number of classes, yielding a real-time solution. We demonstrate an end-to-end production system incorporating chronological and semantic browsing with our detection framework. *Demo video included.*

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Video Logging

General Terms

Computer Vision

Keywords

Video Concept Detection; Concept Ontology; Multi-modal fusion.

1. INTRODUCTION

The emergence of 'shoot to cloud' as a paradigm for broadcast production promises cost efficiencies through the immediate availability of content to downstream studio production. Video shot on-set is streamed from cameras to cloud infrastructure, where web-based studio applications enable users to collaboratively create digital content. Production metadata, 'logged' at the time of capture, is essential to making sense of the deluge of video content acquired during a shoot; downstream production cannot proceed without it. Unfortunately contemporary practice remains reliant upon pen-and-paper for such 'video logging'. Metadata describing the visual content within each shot is manually generated, and often manually re-keyed into digital form stalling an otherwise live pipeline.

Recent advances in the field of Computer Vision, have yielded a step-change in the performance of image classification that for the first time approaches practical requirements for automated metadata annotation [12]. Despite significant progress in tackling diverse datasets of static imagery [12, 6], video classification has been explored only very recently for domain-constrained footage (e.g. datasets of sports video [11]). A key challenge in translating these early successes to diverse, general video 'in the wild', is the consideration of multiple feature modalities in order to discriminate between diverse content classes.

The contribution of this paper is a scalable, real-time framework for video logging that incorporates multiple feature modalities for content classification. Our key technical contributions are:

1. **Class scalability:** Visual concept detection requires a bank of classifiers to be evaluated for each image at test-time, typically either through one-vs-all or one-vs-one support vector machines (SVMs) scaling respectively with $O(c)$ or $O(c^2)$ in the number of classes. Recently the ensemble of exemplar-SVM (EE-SVM) framework [14] has shown greater robustness to content diversity (appearance variability) but at exponential complexity $O(e^n)$ in both the number of classes and training exemplars. Manual video logging relies upon concretely defined ontologies for mark-up, which we leverage in our automated approach to perform hierarchical decision offering a sub-linear test-time complexity of $O(\log c)$ necessary for real-time processing.

2. **Multi-modal fusion:** Dynamic per-class balancing of multiple visual features, comprising primary features (such as SIFT) as well as contextual features such as visual gist, camera motion and foreground-background segmentation to maximise discrimination within each branch of the concept ontology. Prior approaches to video classification have relied upon multi-kernel learning (MKL) or per-element weighting within random decision forests (RDFs). Neither has been shown to scale sub-linearly and the latter typically responds poorly to appearance diversity (a problem compounded when considering multiple feature modalities).

Classification is performed using spatio-temporal features local to keyframes detected within the video stream, and aggregated as a post-process over any temporal interval (typically on a whole-clip basis for logging) using morphological sieving. An interactive visualization enables both chronological and concept-based browsing of the resulting metadata-enriched content.

We demonstrate the proposed system on diverse real-world pre-production ('rushes') footage, contrasting performance against a state-of-the-art baseline using Vector Quantization of dense SIFT features. We demonstrate performance gains both in terms of accuracy and efficiency using our combined hierarchical and multi-modal fusion approach.

2. RELATED WORK

Despite many years of intensive research, semantic video concept detection remains an open Computer Vision challenge. Significant advances have been made in the past decade, driven by competitive benchmarks such as the TRECVid, VideoOlympics and PASCAL challenges where impressive performances have been returned on datasets of increasing size and complexity.

Significant advances in generalised concept detection arrived approximately a decade ago, through the combination of gradient-domain features and codebook quantization referred to as the 'bag of visual words' (BoVW). Initially explored for image similarity (retrieval) applications [18, 7], the BoVW representation was rapidly adopted for concept detection through incorporation of supervised classifiers e.g. support vector machines (SVMs) [8, 17]. Variations on the three core stages of this pipeline (feature extraction, representation, classification) have been extensively explored for image classification. Frequently SIFT or SURF features (over luminance and/or chroma channels) are detected sparsely, or more commonly now sampled densely for the first stage. Various quantization strategies for the second stage have been explored including nearest-neighbour codeword assignment (BoVW) [8], soft-assignment variants of BoVW based on k-NN, sub-quantization of each cluster (VLAD) [2], and encoding of parametric distance to per-codeword Gaussian Mixture Models (Fisher Vector variants) [16]. Third stages of leading techniques almost universally focus on SVMs which for concept detection are multi-class arrangements of various forms e.g. one-vs-all [1]. Cascading hierarchies of SVMs comprising low-level features at leaf nodes, combining into successively higher level conceptual layers of classifiers have been shown to scale well for video classification tasks. Note that such cascaded approaches differ from that of our ontological hierarchy [19]. Although not within the scope of our fully automatic scenario, many user-assistive (semi-interactive) retrieval and classification video systems have also been proposed.

Recently multitudes of simpler classifiers have been combined in novel ways, via randomized decision forests and ensemble of exemplar-SVMs (EE-SVM) [14]. The latter have shown good generalization performance over wide variations in appearance, since each exemplar is effectively a class within detectable concepts effectively super-classes of these. Unfortunately eSVMs exhibit exponential complexity in the number of classifiers wrt. both concept count and training dataset size. Most recently this has resulted in run-time optimization strategies for efficient pruning of eSVMs to reduce the number of classifiers to be tested [15]. Nevertheless these approaches do not reduce complexity and with single image classification times of several seconds an extension to streaming video is not yet realistic. With the advent of ontology structured large-scale image datasets (ImageNet) new opportunities exist to exploit semantic relationships between concepts during detection. Ferrari et al. briefly explored this idea through experiments in hierarchical classification for ImageNet [9], and we follow in this spirit exploring not only ontologically aware classification (for video) but also dynamic weighting of feature modalities in order to distinguish locally occurring concepts within that hierarchy. Significant progress has been made in the last couple years using deep learning to train both classifier and learn optimal features for classification simultaneously. Contemporary approaches using Convolutional Neural Networks (CNNs) have delivered a step change in accuracy for image recognition as end-to-end classifiers [12], although often SVMs are also employed to classify features pulled from the fully connected early stages of the network [6]. However these approaches have yet to be proven on video outside tightly constrained domains e. g. sports footage [11].

A substantial divide exists between performance on research lab datasets versus the class and appearance diversity exhibited by video content encountered 'in the wild'. Barriers to adoption include not only accuracy levels but also speed, with sophisticated classification techniques often requiring substantial training and/or test times that prohibit real-time application. Unfortunately both barriers must be overcome for practical application in a video logging scenario. The volume of content acquired on typical shoot (several days of rushes for a half hour TV drama), coupled with the financial impetus to accelerate downstream production, is currently answerable only via real-time manual annotation as the camera rolls. As digital production moves towards an immediate availability 'shoot to cloud' model, these practical barriers to automated meta-data creation must be overcome.

3. SCALABLE CONCEPT DETECTION

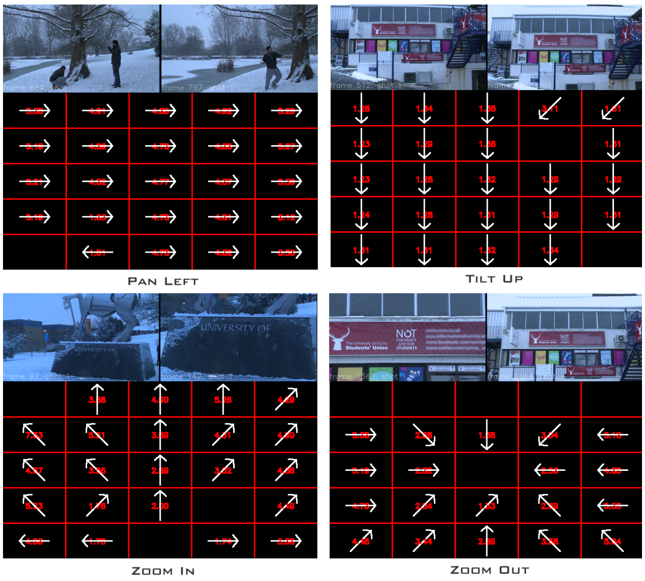
We now describe our scalable system for video concept detection. First we outline the feature extraction process, through which multiple complementary visual features are distilled from ingested footage (Sec. 3.1). We then describe how the hierarchical concept ontology accompanying the footage can be leveraged to both train the system, and perform efficient classification (Sec. 3.2). The mechanism by which the visual features are weighted and combined is then explained (Sec. 3.3). Classification is performed on a per key-frame basis (for K regularly sampled keyframes per clip), but can be aggregated on a clip-level for more convenient browsing in our system. We briefly outline our aggregation strategy to form clip-level classification (Sec. 3.4).

3.1 Feature Extraction

Video is ingested to our system downsampled to QVGA (320×240) resolution prior to extracting the following features on a per key-frame basis.

3.1.1 Segmentation

We first apply motion segmentation to determine foreground and background regions within each key frame. Optical flow vectors [4] are computed densely between the key frame and its previous frame, and global scene motion subtracted in order to produce a camera ego-motion compensated flow estimate. We refer to this flow field as $\mathcal{V}(t)$. Camera motion is approximated by computing the inter-frame homography between the same pair of frames using robust estimation (RANSAC) over the optical flow correspondences. Thresholding $|\mathcal{V}(t)|$ at high and low values yields a fore-


Pan Left

Tilt Up

Zoom In

Zoom Out

Figure 1: Illustrating the spatial grid of average flow vectors used to construct our contextual camera motion descriptor. Discriminative responses are shown for four common camera motions.

ground and background mask for each frame. Morphological closure is performed to remove salt and pepper noise from the masks. Thus we are able to compute the following features over the full-frame (FF), foreground (FG), and background (BG) regions.

3.1.2 *Dense SIFT*

Dense SIFT features are extracted using a three pixel (3px) spacing over four spatial scales, deemed to be optimal over this resolution of footage via prior experimentation [5]. We found that 3px spacing provided 5% accuracy (mAP) boost on our public comparison (Caltech-101) dataset vs. 8px spacing, and 2px spacing provided no benefit and increased the computation cost dramatically. Similarly, a single scale approach was 5% worse than using four SIFT scales (4,6,8,10 over just 4).

Vector quantization is performed using k-means clustering over 10 million features randomly sampled from our dataset, where codebook size of $k = 600$ was found to give best results. A spatial pyramid histogram is then computed as in [13], with L = 2. However, unlike with [13] and under the guidance of [5], no normalisation is used since the frame size is equal and all histograms sum to the same number of extracted features. The BoVW histograms are computed for all three of the segmentation cases (FF, FG, BG) with only SIFT features within the appropriate mask used to build the spatial pyramid histogram.

3.1.3 *Contextual camera motion*

The presence of visual concepts is often correlated with the presence of other contextual information in the scene. For example, certain camera movements are like in the presence of certain semantic concepts. We allow for the learning of these correlations by encoding context in several ways, starting with global camera motion. This is encoded independently for FF, FG and BG regions.

We divide the region into a 5×5 spatial grid. For each grid cell, $\mathcal{V}(t) > \varepsilon$ are quantised into one of eight orientation bins to create a histogram of oriented flow. The histograms from each grid square are then concatenated together to create a 200-dimensional motion histogram, which is then L_2 normalised. All bins are set to zero for grid squares where the vast majority of $\mathcal{V}(t) < \varepsilon$ to prevent noise from skewing the results with unreliable flow orientation. Fig. 1 illustrates the dominant response different camera movements generate within each cells and thus may be discriminated by the motion descriptor.

3.1.4 *Colour Features*

Another valuable form of context is global illumination and colour cast within a frame. We compute a global colour histogram (GCH) in CIELab space, within each of the four spatial quadrants of the frame. The GCH quantises the chroma channels of the space into 16 bins (luminance is disregarded for robustness) resulting in a 128 dimensional global colour descriptor. The GIST descriptor is additionally computed, using the public implementation provided by [10]. Again the descriptors are calculated over the FF, FG and BG regions independently.

3.1.5 *Facial Grid*

A 9-dimensional descriptor was also extracted by counting the number of faces detected in each square of a 3-by-3 spatial grid. Faces are detected using the multi-scale Haar cascade approach due to Viola and Jones [20]. The descriptor is calculated independently over FF, FG and BG regions.

3.2 Hierarchical Path Finding

Multi-class video concept detection is performed via a bank of SVM classifiers χ_c for a concept set $\mathcal{C} = \{c_1, \ldots, c_n\}$, the outputs of which are aggregated to make an overall classification decision $\mathcal{D}_t$ for a multi-modal feature set $\mathcal{F} = \{F_t^1, \ldots, F_t^m\}$ extracted local to key frame I_t. In the commonly used *one-vs-all* classifier arrangement for SVMs, the decision is the maximum probability concept detected over the classifier bank:

$$\mathcal{D}_t(\mathcal{F}) = \underset{c \in \mathcal{C}}{\operatorname{argmax}}\, \chi_c(\mathcal{F}_t). \tag{1}$$

We aim for a decision $\mathcal{D}_t(\mathcal{F}(I_t)) = c, \forall I \in G(c)$ for all frames in ground-truth annotated test set $G(c) \subset I_{\forall t}$ for category c. Classifier χ_c is trained using a similarly annotated training set $T(c)$.

There are two disadvantages with this commonly used framework [5]. First, it scales linearly i. e. $O(|\mathcal{C}|)$. Second, it does not capitalise on ontological relationships within $\mathcal{C}$ i. e. χ_c is trained using positive $\mathcal{F}^+$ and negative $\mathcal{F}^-$ exemplars drawn from $T(c)$ and $T(\mathcal{C} \setminus c)$ respectively. There is opportunity for exploiting these semantic relationships; for example if D_t can be determined with high certainty to be a vehicle then a subsequent classification into categories car, bus, truck, etc. can be made using a classifier more precisely trained over imagery from that sub-branch of the video ontology.

We therefore constrain training according to ontological relationships, where a each concept $c_i^{\uparrow p} \in \mathcal{C}$ has a parent concept $p = c_j$ where c_0 is a dummy concept introduced to indicate the root node (and outcome $\chi_0 = 0$). Classfier χ_i is trained using $\mathcal{F}^+$ drawn from node c_i and descendents $c_j^{\uparrow i}$, $c_k^{\uparrow j}$ and so on recursively — with $\mathcal{F}^-$ drawn from the remainder of T.

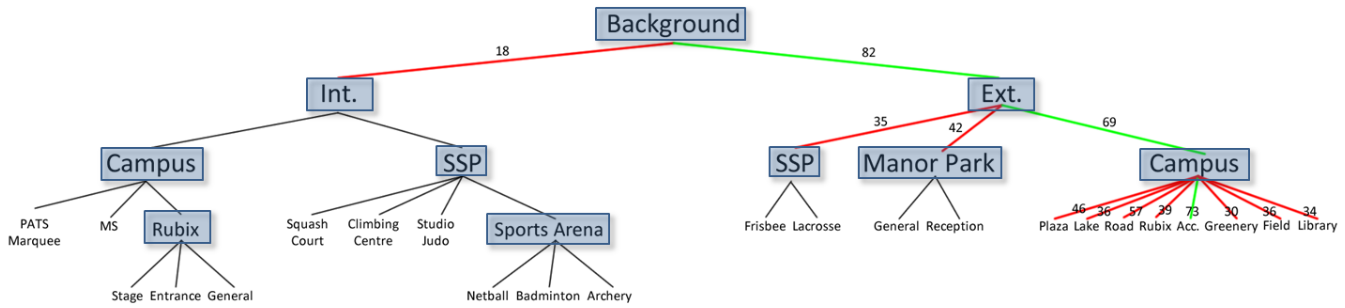


Figure 2: Classification decision for a frame of 'StudentTV' performed using our hierarchical path finding approach. Only 3 SVMs are tested in the best-first traversal versus 23 SVMs for a linear one-vs-all strategy. Values for χ_c computed during traversal are noted beside each node.

At test-time the hierarchy is exploited to reduce the number of classification decisions, presenting a substantial efficiency saving over a one-vs-all strategy to classifying each of K keyframes within lengthy pre-production video clips.

A working set $\mathcal{W}$ maintaining outcomes (*concept-probability* pairs) of classifications at the perimeter of the explored hierarchy is created, initially containing $\{0, \chi_0\}$ for the root node. Classification of a given frame I_t begins by extracting feature set $\mathcal{F}$ (see Sec. 3.1) and measuring $\chi_c(\mathcal{F})$ for c_i^{+p} for $p = 0$. The classification outcomes $\{c, \chi_c + \chi_p\}$ are added to $\mathcal{W}$ and outcomes χ_p removed from $\mathcal{W}$. Exploration of the hierarchy is then advanced one node at a time, each time selecting classifier χ_c' for evaluation and addition to $\mathcal{W}$ where:

$$c' = \underset{c}{\operatorname{argmax}} \mathcal{D}_t(\chi_c(\mathcal{F})). \tag{2}$$

Fig. 2 illustrates the resulting best-first traversal of the ontology for a classification performed for a frame of pre-production footage from 'StudentTV' (Sec. 4.1). Classification halts at the first (most promising) leaf node reached for the quantiative results presented here, but in our user interface (Sec. 4.4) the top few results can be visualised e. g. using all leaves. In the best case $O(\log|\mathcal{C}|)$ classifiers will be tested, and at worst $O(|\mathcal{C}|)$. Note that although prior work has observed closer visual distances between ontological siblings (e. g. over ImageNet [9]) such relationships have not previously been exploited for efficient video classification.

3.3 Multimodal Fusion

Our system adopts a late fusion strategy to combining features across multiple modalities, enabling per-class control over the relative importance of each modality. Consequently χ_c is determined by a weighted combination of linear classifiers $\psi_{c,i}$

$$\chi_c = \frac{1}{m}\sum_{i=1}^{m} \mathcal{D}(\omega_{c,i}\psi_{c,i}). \tag{3}$$

for each h_i we use a linear SVM using the chi-squared kernel mapping with parameter $\lambda = 10^{-3}$ and maximum training iteration count of 5×10^4, using the VLFeat implementation.

Weights $\omega_{c,i}$ are determined on a per-class and are learned over $T(c)$. First, eq. 3 is evaluated for a particular sub-classifier i i. e. with $\omega_{c,i}$ unity and all other weights zero. All training data for concepts at level c or below within the hierarchy are fed through χ_c. Decision score χ_c will be positive or negative for each training frame, and this is determined to be a true positive (TP), or a true (TN) or false (FN) negative according to the training markup for the classifier $\psi_{c,i}$ operating over a specific modality. The discriminatory power of that modality for concept c is estimated via the average of its True Positive Rate (TPR) and Positive Prediction Value (PPV):

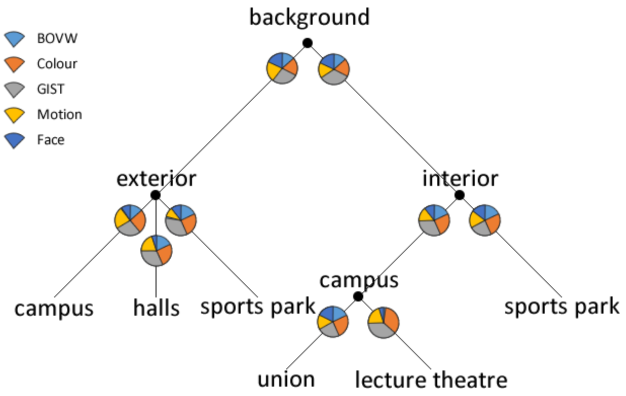


Figure 3: Illustrating the relative weights of the five features dynamically learned on a per-node basis within the hierarchy to perform multi-modal fusion in our system.

$$TPR(c,i) = \frac{TP(c,i)}{TP(c,i)+FN(c,i)}. \tag{4}$$

$$PPV(c,i) = \frac{TP(c,i)}{TP(c,i)+FP(c,i)}. \tag{5}$$

$$\omega(c,i) = \frac{TPR(c,i)+PPV(c,i)}{2}. \tag{6}$$

The weights are normalised such that $\sum_{i=1} m\omega_{c,i} = 1$ and recorded for use in the hierarchical classification.

3.4 Aggregation for clip classification

Classification is performed for each video frame independently, which is important in our pre-production scenario where individual video clips can be up to several minutes long. However for browsing folders ('bins') of video files, and to apply standard benchmarking methodologies, it is necessary to combine frame based decisions to deliver a single representative clip-level classification.

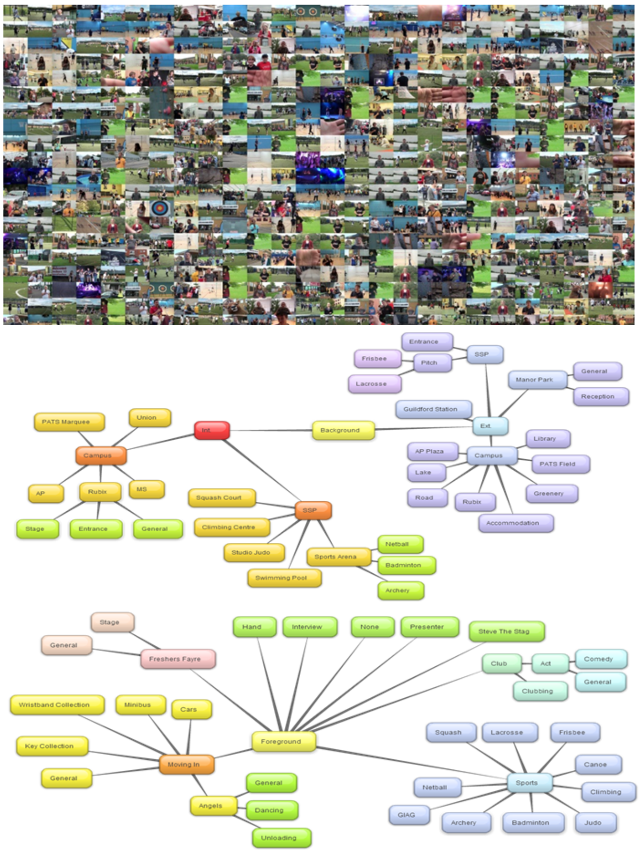


Figure 4: Visual précis (top) of the 'StudentTV' dataset used in our evaluation, comprising 45 concepts (23/22 fore/background respectively) and 1040 video clips of up to 30 minutes duration. The ontologies for background (middle) and foreground (bottom) are illustrated.

Morphological sieving [3] is used to smooth the classification decision signal $\chi_c(\mathcal{F}_t)$ for each concepts independently over time t, preserving large contiguous blocks of detections over time without the loss of high-frequency information at the start and end of such blocks. The sieving is performed over all K key frames of the clip being classified.

4. RESULTS AND DISCUSSION

We evaluate the proposed cloud video classification system over three datasets of varying complexity (subsec. 4.1) reporting performance in terms of both accuracy and efficiency. Table 1 summarises the results obtained. For purposes of future comparison the latter is benchmarked using a single node of Amazon S3 cluster, with an Intel Xeon Processor E5-2670 and 16GB RAM.

All experiments were run with three repetitions of cross-validation over a random train-test split on each dataset, with each split comprising a even ratio of 30 training:test examples for each concept.

4.1 Evaluation Datasets

Large corpora of annotated pre-production video are challenging to obtain; the largest public video datasets for object classification focus primarily on a single domain (e. g. Youtube Sports [11]) which not only lack subject diversity but are not representative of the unfinished and lengthy pre-production clips we address. We have collected and annotated two datasets from real-world broadcast pre-production and post-production to both measure performance and illustrate this distinction.

The primary dataset 'StudentTV' used in this evaluation is a large 1040 video dataset comprising one week of raw pre-production ('rushes') footage shot during Fresher's Week on a University campus location, and comprising both 'A roll' (principal subject matter) and 'B roll' (contextual footage) content. The average duration of a clip is 40 seconds, however some clips are up to 30 minutes long. A visual précis of this footage is presented in Fig. 4 alongside a concept ontology accompanying the dataset. The ontology is divided into foreground and background branches, which contain a total of 45 (23 and 22 respectively) leaf-node concepts within a hierarchy up to 5 concepts deep and comprising 54 decision nodes. To enable detailed evaluation we examine foreground and background performance separately.

Two smaller datasets are also compared against, for the purpose of additionally characterising performance of the hierarchical and multi-modal fusion aspects of the system. First, a smaller video dataset 'Hospital' is included for purposes of scalability comparison with StudentTV. The dataset comprises 98 videos of post-production footage of a TV hospital drama, each clip being of average duration 5 seconds. The concept ontology for this footage is flat (i. e. depth 1) comprising only 4 concepts. Second, for the purposes of comparing multi-modal fusion over a public benchmark we analyse the Caltech-101 image dataset, which is supplied with 101 concepts which we organised into an ontology with hierarchy 2 levels deep.

4.2 Classification Accuracy

Fig. 5a summarises the performance in terms of accuracy (mean average precision) over all three datasets, both for the full system and for restricted configurations of the system with various components disabled in order to show their contribution. In all cases the evaluation has been run at the clip level i. e. using the aggregation post-process of subsec.3.4 to integrate the individual classifications made for K key frames within each clip (we use $K = 5$ for all experimental values reported).

We experiment with combinations of the following cases of restriction: 1) disabling the hierarchical path finding, effectively flattening the ontology to a hierarchy a single layer deep (FLAT); 2) disabling the dynamic weight calculation for the multi-modal fusion, and adopting a uniform weighting instead (UNIFORM); 3) disabling the foreground/background segmentation (NOSEG) using features from the FF region (subsec. 3.1.1) rather than the FG or BG regions independently; 4) disabling all features except for SIFT (SIFTONLY). The combination of these restrictions yield our baseline SIFT / Vector Quantization system for the purposes of comparison. The combination of these restrictions, excluding FLAT, represent hierarchical path finding over the BASELINE.

4.2.1 Accuracy over 'StudentTV'

The highest performing configuration of the system over 'StudentTV' was the full system without any segmentation pre-process (StudentTV:NOSEG), resulting in overall 58.8% MAP (60.1% and 57.7% respectively for background and foreground concept groups). This represents a significant improvement (close to doubling of perfor-

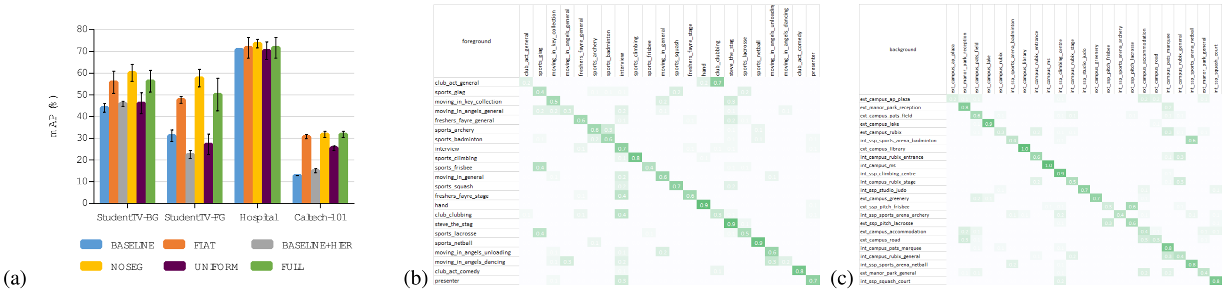


Figure 5: (a) Accuracy (mAP) of the proposed system over three datasets: 'StudentTV', 'Hospital', 'Caltech-101'. Refer to Sec. 4.2 for meaning of abbreviations which relate to system configurations. Confusion matrices summarising per-class performance of the proposed system over the foreground and background classes of the 'StudentTV' dataset.

Test Configuration	StudentTV-BG		StudentTV-FG		Hospital		Caltech-101	
	Accuracy (%)	Nodes	Accuracy (%)	Nodes	Accuracy (%)	Nodes	Accuracy (%)	Nodes
BASELINE	44.0 ± 1.9	23.0	31.1 ± 2.7	22.0	70.8 ± 0.0	4.0	12.8 ± 0.1	101.0
BASELINE+HIER	45.9 ± 1.2	9.3	22.5 ± 1.8	12.0	70.8 ± 0.0	4.0	15.0 ± 0.9	42.3
FLAT	55.8 ± 5.1	23.0	47.7 ± 1.5	22.0	71.7 ± 4.7	4.0	30.7 ± 1.0	101.0
NOSEG	**60.1 ± 3.9**	**10.0**	**57.7 ± 4.0**	**12.0**	**73.6 ± 2.0**	**4.0**	**31.8 ± 1.5**	**47.0**
UNIFORM	46.1 ± 4.8	9.7	27.2 ± 4.7	11.7	70.3 ± 4.1	4.0	25.5 ± 0.9	45.3
FULL	56.3 ± 5.0	10.0	50.1 ± 7.5	12.0	71.7 ± 4.7	4.0	**31.8 ± 1.5**	**47.0**

Table 1: Accuracy (mAP) of the proposed system over three datasets: 'StudentTV', 'Hospital', 'Caltech-101'. The number of classification nodes visited within the hiearchy is also reported. Refer to Sec. 4.2 for meaning of abbreviations which relate to system configurations.

mance) over the SIFT/BoVW baseline implementation (StudentTV:BASELINE) which scored 37.1% (44.0% and 31.1% respectively). The confusion matrices presented in Fig. 5 (b,c) characterise the performance of the system on a per category basis.

We evaluated the impact of the hierarchical path finding on accuracy, by flattening the ontology to a hierarchy a single layer deep (StudentTV:FLAT). Compared to the full system, performance of StudentTV:FLAT dropped slightly by 2.4% and 0.5% for background and foreground respectively, demonstrating a small accuracy benefit through use of the hierarchical scheme. Although the main purpose of the hierarchical search was to reduce number of SVM comparisons, rather than boost accuracy, the latter may have arisen due to the reduced scope for distraction caused by the reduction. The more focused training sets created through consideration of the hierarchy when identifying per-class positive and negative examples may also deliver some benefit.

Disabling the dynamic weighting and running the system with uniform priority across all features for all classes (StudentTV: UNIFORM) resulted in overall MAP of 36.9% (46.1% and 27.2%, for foreground and background respectively) representing a significant drop in performance versus the dynamically weighted results and similar performance to BASELINE, within tolerance. This shows that simply adding more features to a classification system alone does not appreciable improve the performance, rather careful attention must be paid to the fusion strategy. The result clearly illustrates the benefit of setting per-class feature weightings automatically at training time, rather than prescribing a single global weighting of features for the dataset.

The proposed system shows significant accuracy benefit versus the baseline for the most challenging dataset, StudentTV. Whilst accuracy is increased via our two core contributions (hierarchy and multi-modal fusion) the proposed use of foreground/background segmentation based on motion cues underperforms on average versus StudentTV:NOSEG. Despite the separation of ontologies and use of FG/BG regions over FF did not, as might intuitively be expected, produce higher results (despite features for separate FG and BG regions being available). Fig. 6 (top) visualises the relative per-class performance of StudentTV vs. StudentTV:NOSEG, which combined with the larger standard deviation of the former, indicates that some classes benefit from significant performance increases using segmentation (especially in the background ontology). Many of the best performing classes under segmentation contained distinct, fast moving objects (e. g. sports clips) which appear to fare well under the motion segmentation technique used (Fig. 6, bottom).

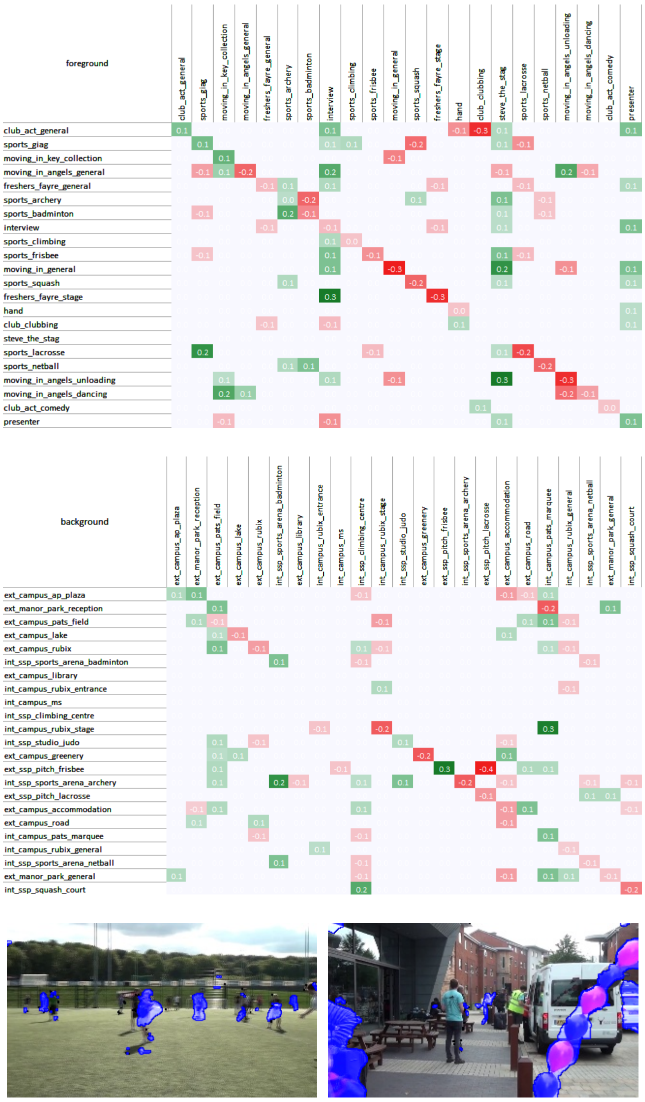
4.2.2 *Accuracy over 'Hospital' and 'Caltech-101'*

Accuracy over smaller datasets was evaluated to provide an indication as to the scalability of the system ('StudentTV' vs. 'Hospital') over production footage, and against a standard image classification dataset ('Caltech-101'). For the latter, motion features could not be included in the multimodal fusion, though a performance boost almost doubling the mAP is observed when the remainder of the feature bank is included and dynamically weight. Where few concepts exist and the ontology is flat ('Hospital') or shallow and simplistic ('Caltech-101') there is little benefit in our hierarchical path finding approach — although, as with 'StudentTV', there is some minor benefit over the BASELINE of a few percent.

4.3 Classification Efficiency

4.3.1 *Performance over 'StudentTV'*

For the StudentTV dataset the average time taken to classify a video clip for the best performing system configuration (StudentTV:NOSEG) was 22.5 seconds (23.2 and 21.6 seconds respectively for the background and foreground ontologies). Given an average clip length of 30 seconds, the classification speed averages out at slightly bet-



foreground	club_act_general	sports_giag	moving_in_key_collection	moving_in_angels_general	freshers_fayre_general	sports_archery	sports_badminton	interview	sports_climbing	sports_frisbee	moving_in_general	sports_squash	freshers_fayre_stage	hand	club_clubbing	steve_the_stag	sports_lacrosse	sports_netball	moving_in_angels_unloading	moving_in_angels_dancing	club_act_comedy	presenter
club_act_general	0.1							0.1						-0.1	-0.3	0.1						0.1
sports_giag		0.1						0.1	0.1			-0.2				0.1	-0.1					
moving_in_key_collection			0.1								-0.1											
moving_in_angels_general		-0.1	0.1	-0.2				0.2								-0.1			0.2	-0.1		
freshers_fayre_general					-0.1	0.1		0.1					-0.1				-0.1					0.1
sports_archery						0.0	-0.2					0.1				0.1		-0.1				
sports_badminton		-0.1				0.2	-0.1									0.1		-0.1				
interview					-0.1			-0.1					-0.1			0.1						0.1
sports_climbing								0.1	0.0													
sports_frisbee		-0.1						0.1		-0.1						0.1	-0.1					
moving_in_general								0.1			-0.3					0.2			-0.1			0.1
sports_squash						0.1						-0.2				0.1						0.1
freshers_fayre_stage								0.3					-0.3									
hand														0.0								0.1
club_clubbing					-0.1			-0.1						0.1								0.1
steve_the_stag																						
sports_lacrosse		0.2								-0.1						0.1	-0.2					
sports_netball						0.1	0.1											-0.2				
moving_in_angels_unloading			0.1					0.1			-0.1					0.3			-0.3			
moving_in_angels_dancing			0.2	0.1															-0.2	-0.1		
club_act_comedy															0.1						0.0	
presenter			-0.1					-0.1								0.1						0.1

background	ext_campus_ap_plaza	ext_manor_park_reception	ext_campus_pats_field	ext_campus_lake	ext_campus_rubix	int_ssp_sports_arena_badminton	ext_campus_library	int_campus_rubix_entrance	int_campus_ms	int_ssp_climbing_centre	int_campus_rubix_stage	int_ssp_studio_judo	ext_campus_greenery	ext_ssp_pitch_frisbee	int_ssp_sports_arena_archery	ext_ssp_pitch_lacrosse	ext_campus_accommodation	ext_campus_road	int_campus_pats_marquee	int_campus_rubix_general	int_ssp_sports_arena_netball	ext_manor_park_general	int_ssp_squash_court
ext_campus_ap_plaza	0.1	0.1								-0.1							-0.1	-0.1	0.1				
ext_manor_park_reception			0.1																-0.2			0.1	
ext_campus_pats_field		0.1	-0.1								-0.1							0.1	0.1	-0.1			
ext_campus_lake			0.1	-0.1													0.1						
ext_campus_rubix			0.1		-0.1					0.1	-0.1								0.1	-0.1			
int_ssp_sports_arena_badminton						0.1				-0.1											-0.1		
ext_campus_library																							
int_campus_rubix_entrance											0.1									-0.1			
int_campus_ms																							
int_ssp_climbing_centre																							
int_campus_rubix_stage								-0.1			-0.2								0.3				
int_ssp_studio_judo			0.1		-0.1							0.1					-0.1						
ext_campus_greenery			0.1	0.1									-0.2				0.1						
ext_ssp_pitch_frisbee			0.1						-0.1					0.3		-0.4		0.1	0.1				
int_ssp_sports_arena_archery			0.1			0.2	-0.1			0.1		0.1			-0.2		-0.1				-0.1		-0.1
ext_ssp_pitch_lacrosse																-0.1					0.1	0.1	
ext_campus_accommodation		-0.1	0.1							0.1							-0.1	0.1					-0.1
ext_campus_road		0.1			0.1												-0.1						
int_campus_pats_marquee					-0.1					-0.1									0.1				
int_campus_rubix_general								0.1												-0.1			
int_ssp_sports_arena_netball						0.1				-0.1											-0.1		
ext_manor_park_general	0.1									-0.1							-0.1		0.1	0.1		-0.1	
int_ssp_squash_court										0.2													-0.2

Figure 6: Top, Middle: Confusion matrices showing relative performance of using segmentation or not on a per-class basis. Bottom: Motion-based foreground segmentations of good and poor performing classes.

ter than real-time (but note that classifications are aggregated from key-frames only). The performance of the full system including segmentation drops to slightly lower than real-time at 36.4 seconds due to the computational overhead of the motion-segmentation algorithm.

Table 1 summarises the number of SVM comparisons made on average per classification, which gives best indications on the scalability of the system with the number of classes. On average, approximately half the number of node evaluations must be made in order to reach a classification decision when using the hierarchical path finding. In FLAT cases (including the BoVW BASELINE) all nodes must be investigated (under the one-vs-all framework) to reach a classification decision. This represents a doubling in performance for our test scenario. We note that the test-time complexity of our hierarchical strategyis $O(logn)$ in the best case and $O(n)$ in the most pessimistic (in terms of the number of node evaluations scaling with number of classes). We would anticipate the benefit of this technique to tend to well below half for significantly larger ontologies, or ontologies deeper than that accompanying 'StudentTV'.

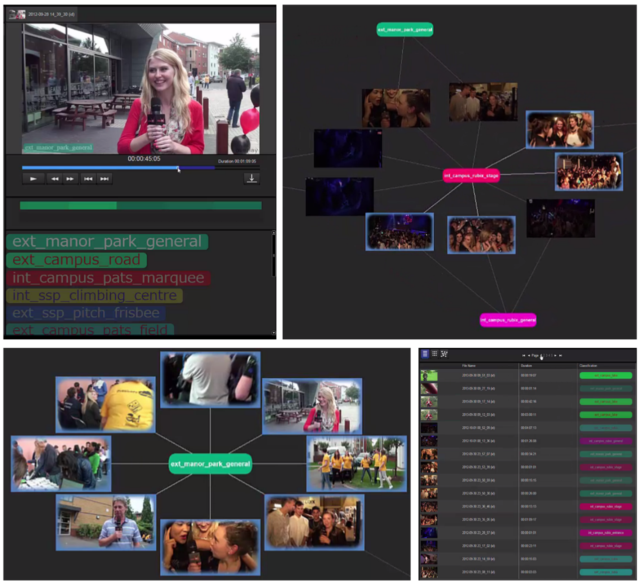


Figure 7: Screens from the prototype commercial shoot-to-cloud system into which our proposed algorithm was integrated. Three forms of visualisation may be used to view classification results for ingested content: 1) Graph 'mind-map' view showing clips sharing concepts; 2) List view; 3) Colour heat-map view with accompanying tag cloud, showing per keyframe classification on a timeline that enables seeking for concepts over longer clips of 'rushes' pre-production footage.

4.3.2 Performance over 'Hospital' and 'Caltech-101'

Similar to 'StudentTV' the hierarchical path finding approach results in around one half as many node evaluations for 'Caltech-101' despite a simpler ontology of only 2 concepts deep. Again, minor improvements in accuracy are observed; the hierarchical path finding is successful even though the hierarchy is shallow. No performance benefit is seen without a hierarchical ontology ('Hospital') as the system degenerates to a 'one-vs-all' comparison, although accuracy benefits are present through fusion of the multi-modal features.

4.4 Content Visualisation

Our classification algorithm was integrated into a prototype commerical 'shoot to cloud' infrastructure on Amazon S3. Fig. 7 illustrates three forms of content browsing interface available via web interface. Three types of visualization are available, drawing upon the metadata tags generated by our automatic video concept detection technique. First, a mind-map enabling clips to be navigated through semantic relationships established by shared concepts. Second, a classical list view showing the most likely classification of each clip in the clip bin. Third, a colour based heatmap and tag cloud showing the most probable classifications on a per-keyframe basis over time. The latter enables rapid seeking for target concepts within pre-production rushes footage, which can be relatively long (up to 30 minutes per clip for 'StudentTV') relative to the 1-2 second shots typically handled by research systems that focus upon post-production content.

5. CONCLUSION

We have presented a novel video concept detection system integrated within a prototype shoot-to-cloud video infrastructure, and evaluated its performance over a challenging > 1000 video dataset of pre-production ('rushes') footage. Two additional smaller datasets were compared against for reference. Two novel technical contributions have been presented:

First, a hierarchical path finding algorithm that significantly reduces the number of classification decisions that must be made for each frame of content. By adopting an efficient best-first traversal of the concept ontology the number of classification decisions in our system scales $O(\log n)$ with the number of concepts, and linearly with the number of frames processed. We demonstrate ingestion and classification times approximately equal to the length of the clip, enabling real-time data rates on average and making our system practical for shoot-to-cloud scenarios. Moreover, the training of classifiers using ontologically local concepts appears to yield a secondary benefit through a minor improvement in classification accuracy.

Second, the fusion of multiple feature modalities (e.g. colour, texture, motion, face presence) is proposed using dynamically learned importance weightings for each modality at each node of the classification hierarchy. Increases in performance are shown using a fusion of multiple modalities over single feature (SIFT) classification, and these are most significant for the large, diverse 'StudentTV' dataset where greatest benefit is also shown for the dynamic learning of weights on these features.

We have also investigated the possibility of using motion segmentation to extract foreground and background regions from the video, and to classify foreground and background concepts independently within those regions. Whilst a several classes showed improvements in accuracy via this technique, the overall average performance dropped slightly with larger standard deviation. Inspection of the automatic segmentation masks on best and worst performing classes indicates that performance is correlated to segmentation quality, which given the reliance upon motion in this case produced fragmented regions for near-stationary objects. A clear direction for future work is to explore alternative appearance-based segmentation algorithms to overcome this issue.

Acknowledgments

The first author was supported by an EPSRC Industrial Case studentship co-sponsored by Sony Europe / Broadcast Professional Research Labs.

6. REFERENCES

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] R. Arandjelović and A. Zisserman. All about VLAD. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2013.

[3] A. Bangham, K. Moravec, R. Harvey, and M. Fisher. Scale-space trees and applications as filters for stereo vision and image retrieval. In *Proc. British Machine Vision Conference (BMVC)*, 1999.

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 25–36, 2004.

[5] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. British Machine Vision Conference (BMVC)*, 2011.

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. British Machine Vision Conference (BMVC)*, 2014.

[7] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, 2007.

[8] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *Proc. European Conf. on Computer Vision (ECCV)*, 2004.

[9] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 1777–1784, 2011.

[10] M. Douze, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *Proc. Intl. Conf. on Image and Video Retreival (CIVR)*, 2009.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105. 2012.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[14] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, pages 89–96, 2011.

[15] D. Modolo, A. Vezhnevets, O. Russakovsky, and V. Ferrari. Large-scale video classification with convolutional neural networks. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 143–156, 2010.

[17] J. Sivic, F. Schaffalitzky, and A. Zisserman. Efficient object retrieval from videos. In *European Signal Processing Conference*, 2004.

[18] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, volume 2, pages 1470–1477, 2003.

[19] C. G. M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, January 2005.

[20] P. Viola and M. Jones. Robust real-time object detection. *Intl. Journal Computer Vision (IJCV)*, 57(2):137–154, 2004.