

Concept Detection in Multimedia Web Resources about Home Made Explosives

George Kalpakis¹, Theodora Tsikrika¹, Foteini Markatopoulou^{1,2}, Nikiforos Pittaras¹, Stefanos Vrochidis¹,
Vasileios Mezaris¹, Ioannis Patras², Ioannis Kompatsiaris¹

¹Information Technologies Institute, CERTH
Thessaloniki, Greece

{kalpakis, theodora.tsikrika, markatopoulou, npittaras, stefanos, bmezaris, ikom}@iti.gr

²Queen Mary University of London
London, UK

i.patras@qmul.ac.uk

Abstract— This work investigates the effectiveness of a state-of-the-art concept detection framework for the automatic classification of multimedia content, namely images and videos, embedded in publicly available Web resources containing recipes for the synthesis of Home Made Explosives (HMEs), to a set of predefined semantic concepts relevant to the HME domain. The concept detection framework employs advanced methods for video (shot) segmentation, visual feature extraction (using SIFT, SURF, and their variations), and classification based on machine learning techniques (logistic regression). The evaluation experiments are performed using an annotated collection of multimedia HME content discovered on the Web, and a set of concepts, which emerged both from an empirical study, and were also provided by domain experts and interested stakeholders, including Law Enforcement Agencies personnel. The experiments demonstrate the satisfactory performance of our framework, which in turn indicates the significant potential of the adopted approaches on the HME domain.

Keywords— concept detection; concept-based multimedia retrieval; visual feature extraction; home made explosives

I. INTRODUCTION

The large number of terrorist attacks that have taken place around the world over the past 20 years, along with the technological advancements in the field of multimedia analysis and retrieval, have drawn attention to the discovery, analysis, and understanding of multimedia content that possibly depicts terrorist or criminal activities. To enable the classification of multimedia content as relevant to the security domain and to support end-user queries for the retrieval of such content, we need to develop technologies that are able to automatically detect specific security-related objects in such content. In this context, this work focusses on the analysis of multimedia content, namely images and videos, embedded in publicly available Web resources containing recipes for the synthesis of Home Made Explosives (HMEs), and investigates the effectiveness of detecting specific HME-related objects in such content, so as to classify it as relevant to the HME domain and to also support its retrieval with respect to the detected objects.

This would be particularly useful to law enforcement and security agencies in their quest to better understand and thus counter the potential threats arising from the subversive use of this HME-related information, e.g. in acts of terrorism or serious and organised crime. In

particular, the ability to automatically classify online multimedia content as relevant to the HME domain and detect important semantic information within allows for increased efficiency in identifying the chemical materials and everyday household goods needed for synthesising HMEs and thus better monitor the purchase of such items or materials. Moreover, readily available access to information regarding the visual appearance of materials used for building HMEs could support police officers towards a fast identification of chemical substances found on the scene by visual inspection and thus enable them to take appropriate measures based on their potential hazard.

To this end, a state-of-the-art multimedia concept detection framework for the automatic classification of visual content to a set of predefined semantic concepts is employed. The framework relies on recently developed techniques for extracting low-level visual features from videos and images, so as to relate them to higher-level semantic concepts with meaningful and comprehensible interpretation for end users; specifically, the popular SIFT [1] and SURF [2] features, and their variations [3][4][5] are used. The classification process is based on machine learning techniques and specifically on logistic regression, one of the most prominent methods applied by multimedia content classifiers. The effectiveness of this concept detection framework is evaluated on a dataset of images and videos embedded in HME Web resources discovered either manually by domain experts and interested stakeholders, such as Law Enforcement Agencies (LEAs) personnel, or automatically through focussed crawling [6] and search engine querying; the dataset includes videos found on the YouTube social multimedia sharing platform (<http://www.youtube.com>), as well as images accompanying the core textual content of HME websites. The concepts selected for evaluation were either specified by the aforementioned end-users or were derived based on the findings of an empirical study designed to identify the main characteristics of HME Web resources.

The main contribution of this work is the investigation of the effectiveness of a state-of-the-art concept detection framework on the HME domain. To the best of our knowledge, this is the first attempt to automatically analyse visual content related to this specific domain. The remainder of this paper is structured as follows: Section II reviews related work in the field of multimedia content-based analysis and retrieval, along with multimedia concept detection efforts for security purposes. Section III presents the multimedia concept detection framework.

Section IV introduces the HME-related semantic concepts selected for evaluation. Section V describes the experimental set-up of the evaluation procedure. Section VI presents the results of the evaluation experiments. Section VII discusses our conclusions.

II. RELATED WORK

Concept detection on images and/or videos is performed by taking advantage of the features extracted mainly from their visual content; textual and/or audio features may also be employed, but are not considered in this work. The goal is to associate these low-level visual features to higher-level semantic concepts, i.e. objective linguistic descriptions representing entities that can be observed in visual content [7]. This section first reviews state-of-the-art multimedia concept detection approaches and then also discusses multimedia content-based analysis and retrieval techniques for security purposes.

A. Visual concept detection

A typical concept detection system consists of the following sequential steps:

- 1) **Video decoding:** In case the input corresponds to a video file, a series of specific video frames are extracted for further processing.
- 2) **Feature extraction and representation:** This step involves techniques for extracting a set of descriptors, capable of effectively characterising the visual content (images/video frames), by exploiting global or local information (visual features).
- 3) **Classification:** Based on the extracted visual features, this step is responsible for first building and then applying the trained models that classify multimedia content to a set of predefined semantic concepts.

The following sections provide a brief description of the state-of-the-art methods for each of these three steps.

1) Video decoding

This process, known as video segmentation, aims at reliably representing the visual content by extracting the most representative video frames (referred to as *keyframes*); to this end, one of three approaches is typically applied. The first approach performs video segmentation on a *predefined timestamp* basis where the video content is partitioned in keyframes based on a strict time rate. This time rate is user-defined either intuitively or by taking into account the desirable number of the extracted keyframes for representing the video. The second approach, known as *shot segmentation*, aims at dividing each video into continuous interrelated structural segments corresponding to successive frames, called shots. Shot boundaries can be defined by taking into account different types of transitions (either abrupt or gradual ones) that may occur between successive video frames [8][9][10][11][12]. Finally, the third approach, known as *scene segmentation*, results in extracting scenes, which as opposed to shots, are higher-level temporal fragments, giving a broader description of the story-telling parts of the video by representing a single event or a series of events having semantically coherent content. Scene segmentation can be performed by applying graph-based techniques [13], approaches taking advantage of inter-shot similarity [14], methods based on clustering algorithms [15], or statistical techniques [16].

2) Feature extraction and representation

Numerous research efforts have been conducted for describing the visual content of images/video frames in an effective and reliable manner by using global or/and local descriptors. In large-scale video concept detection, local features are typically employed. The most prominent and widely-used local descriptors that aim to visually represent interest points include the Scale-Invariant Feature Transform (SIFT) descriptor [1], and its extensions HSV-SIFT [17], HUE-SIFT [18], opponent-SIFT, rg-SIFT, C-SIFT, and RGB-SIFT [3], as well as the Speeded Up Robust Features (SURF) descriptor [2], along with its variations, such as dense-SURF [19]. Their popularity is due to their invariant properties with respect to intensity, colour, scale and rotation [20][3], with SURF extraction being less computationally demanding [2].

For visual concept detection, local descriptors extracted from different patches of each image/video frame are aggregated into a global image representation; this process is referred to as *feature encoding*. The “bag-of-words” (BoW) representation [21] has been the most popular method in the last years. This method forms a visual vocabulary by means of a clustering procedure which groups similar points of interest in clusters, and considers each cluster as a visual word of this vocabulary. However, several studies [22][23] have indicated that the BoW method, despite its popularity, does not constitute the most efficient approach compared to other more recent methods, such as the Fisher Vector (FV) [24] and the Vector of Locally Aggregated Descriptors (VLAD) [25]. Actually, VLAD is a fast approximation of FV, demonstrating slightly inferior performance; however it is computationally faster and more practical in use [25].

3) Classification

Several research efforts have focussed on the development of effective and efficient classifiers in the field of visual concept detection. One of the most prominent techniques is based on Support Vectors Machines (SVMs) that construct decision planes in multidimensional space serving as the boundary between instances of different classes [26]. Another state-of-the-art approach demonstrating good performance involves relevance feedback algorithms based on logistic regression models [27]. Moreover, research has shown that deep hierarchical neural networks, known as convolutional neural networks (CNNs), constitute a good solution for recognising visual patterns directly from image pixels [28], demonstrating high tolerance to distortion [29], and hence, they can be used for classification purposes. Other efforts have focussed on implementing Gaussian Mixture Models (GMMs), where the classes are modeled based on a multidimensional Gaussian distribution [30]. Finally, researchers have also applied Hidden Markov Models (HMM) for video analysis and classification [31][32].

B. Multimedia content-based analysis and retrieval for security applications

The terrorist attacks and tragic events that took place around the world the past years (and especially after 9/11) have attracted much more attention towards the analysis and understanding of visual content for discovering plausible terrorist actions and criminal activities, including

antisocial behaviour, such as fights, vandalisms, breaking and entering into shop windows, etc. Several research efforts have been conducted for detecting related concepts on video content originating either from public or private surveillance systems and Closed Circuit Television (CCTV) networks, or from online multimedia content sharing platforms, such as YouTube.

One of the early efforts on this domain attempted to recognise vandalism events occurring in a Brussels Metro Station based on CCTV video stream by exploiting semantic information on static objects and interesting areas [33]. More recent research has focussed on detecting people abandoning unattended objects representing possible threats in crowded environments by recognising four relevant sub-events that characterise this activity [34]. Furthermore, another approach has been directed to recognising protected zone violation and vandalism events based on object detection, object localisation and human movement detection on visual data [35]. Moreover, research efforts have been conducted in the field of facilitating law authorities' suspect search on video surveillance data by detecting persons with specific facial fine-grained characteristics [36]. Finally, a method aiming at identifying extremist videos on YouTube based solely on textual user-generated data (e.g. comments, descriptions etc.) for each video has shown good performance on classifying the relevant videos [37]. This work differs from existing research, as it relates to the HME domain, which, to the best of our knowledge, has not yet been previously investigated in a focussed manner, and also analyses the visual content using recently developed state-of-the-art approaches for detecting HME-related concepts.

III. HME CONCEPT DETECTION FRAMEWORK

This section describes the HME multimedia concept detection framework components and the respective implemented approaches depicted in Figure 1: the visual content processing, the feature extraction, the classification, and the fusion components.

A. Visual Content Processing

The visual content processing component involves the video decoding and the image/video frame resizing. Regarding the video decoding, the segmentation of a video into shots is performed using a variation of the algorithm introduced in [11]. The detection of shot boundaries is based on the assessment of the visual similarity between neighbouring frames of the video. For this, the visual content of each frame is represented by computing a set of local (i.e. the ORB descriptors

proposed in [38]) and global (i.e. HSV color histograms) descriptors, thus allowing the detailed matching of a pair of frames and the effective detection of their differences both in color distribution and at a more fine-grained structural level. Then, shot transitions are detected by quantifying the visual similarity between successive or neighbouring frames of the video, and comparing it against experimentally specified thresholds that indicate the existence of abrupt and gradual shot transitions. Finally, the overall detection efficiency of the algorithm is enhanced by using dedicated detectors for the identification of dissolves and wipes, and for filtering false alarms. After performing the video segmentation process, one keyframe from each video shot is kept for subsequent analysis. The resulting video shots together with the images in the collection are resized so as not to exceed a desirable maximum dimension threshold. This procedure is executed strictly for computational reasons, since the adopted feature extraction approaches are scale-invariant (see next section). Here, this threshold was set to 320x240 pixels, which is a good compromise for preserving the necessary information on one hand, and performing feature extraction in a less time and resource-consuming fashion on the other.

B. Feature Extraction

The feature extraction component of the framework exploits local features based both on the popular SIFT descriptor along with its colour-based variations (i.e. RGB-SIFT and opponent-SIFT), and on the widely-used and efficient SURF descriptor and its recently introduced colour-based variations (i.e. RGB-SURF and opponent-SURF) [4][5]. SIFT performs scale-invariant feature extraction and is capable of reliably detecting objects, even among clutter and under partial occlusion. SIFT features, called keypoints, constitute circular orientated image regions and are invariant to image scaling, rotation and viewport change, partially invariant to illumination alterations and robust against geometrical distortion. Likewise, SURF is a powerful scale-invariant local feature detector and descriptor, partially inspired by the SIFT descriptor. It is popular for its high performance and its standard version is several times faster than SIFT. It is well-known for its robustness against several different image transformations.

The feature extraction stage results in creating the following vectors:

- for the standard (grayscale) SIFT and SURF descriptors, vectors with dimensions $128 \times \langle \text{number of keypoints} \rangle$ are created, and

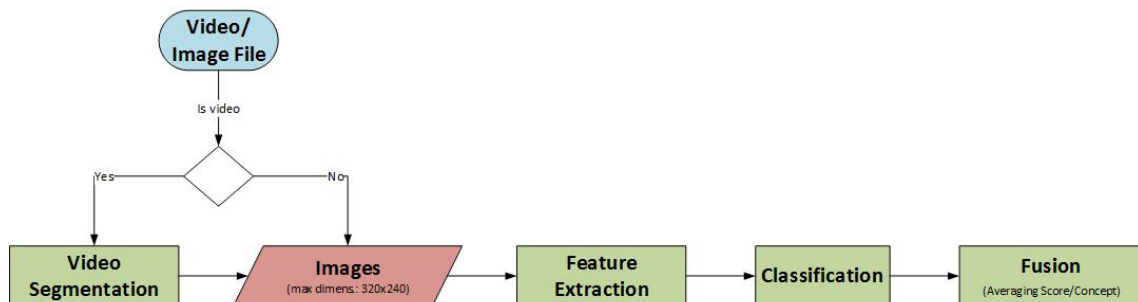


Figure 1. HME multimedia concept detection framework

- for the RGB-SIFT, opponent-SIFT, RGB-SURF, and opponent-SURF (colour-based) descriptors, vectors with dimensions $3 \times 128 \times \langle \text{number of keypoints} \rangle$ are created.

Feature encoding is realised using VLAD, which on one hand is capable of reliably representing an image by aggregating local descriptors in vector space, and on the other allows for remarkable dimensionality reduction by performing principal component analysis (PCA), without impacting its accuracy. The approach adopted is described in detail in [4]. Here, all the local descriptors (SIFT, SURF, and their variations) are compacted to 80 dimensions using PCA and are aggregated using the VLAD encoding. The result of the above process is a VLAD vector of 163840 elements for SIFT or SURF. The VLAD vectors are then compressed into 4000-element vectors by applying a modification of the random projection matrix. These reduced VLAD vectors serve as input to the classification step.

C. Classification

The approach adopted for the classification process is based on logistic regression where the core idea is that a linear model is trained using logistic regression for learning associations between visual characteristics and semantic concepts, based on a set of ground-truth annotated images/video frames serving as training data (where each image/video frame is characterised either as relevant to a higher-level concept or not), by exploiting a number of low-level visual features extracted from each. The trained model is able to classify an unlabelled image/video frame by estimating its confidence score of being relevant to the concept under consideration. The performance of a trained model is assessed in terms of its prediction ability by exploiting a different set of annotated images/video frames, serving as testing data.

Given that we use six descriptors (SIFT and SURF descriptors, and their variations), six distinct classifiers are trained for each concept taking into account the images/video frames belonging in the training set. The classifiers' effectiveness is evaluated on the basis of the testing images/video frames, producing prediction scores in respect to every examined concept for every testing image/video frame and each trained model.

D. Fusion

The final component of our concept detection framework performs late fusion in the semantic space. This entails that at this stage an average prediction score per concept is produced, taking into account the confidence scores calculated by the six classifiers.

IV. HME-RELATED CONCEPTS

The principal objective of this work is to detect higher-level visual concepts in HME multimedia (both images and videos), so as to support information retrieval with respect to the visual content in response to text-based user queries. To this end, and given the lack of available concept lexicons for the HME domain, a set of HME-related concepts was determined based on end-user requirements and also on the findings of an empirical study conducted for identifying the main characteristics of HME visual content.

A. User-specified HME-related concepts

Defining the concepts of interest for the HME domain is of vital significance for the effectiveness of HME multimedia content analysis and retrieval. In this context, we considered that the actual employment of HME domain experts and interested stakeholders, such as LEA personnel, proposing a set of semantic concepts based on their needs and requirements (e.g. for retrieving multimedia content based on text-based queries related to HME concepts), is one of the most effective ways for obtaining a useful and well-reasoned initial concept list. Their prevalent advantage is their profound knowledge of the domain's core characteristics and terminology. Within this context, it is readily understood that their contribution will result in forming a concise initial HME concept list.

To this end, a questionnaire was addressed to the HME domain experts and personnel of two European LEAs and one European Defence Institute. The questionnaire was related to the preferable semantic concepts to be identified in HME-relevant multimedia content, so as to support retrieval in response to text-based queries.

The responses to these questionnaires indicate that the directives of domain experts and interested stakeholders are mainly associated with the characteristics of the materials used in recipes for synthesising HMEs. Such materials can be found in the form of liquids or solids, such as powder, crystals, granular mixtures, and plasticised or slurry structures, in different colour variations and states (e.g. evaporating or fuming liquids). In particular, concepts such as "clear liquid", "white powder", "yellow crystals", "white granules" or "yellow fuming liquid", are considered important for enabling the retrieval of multimedia content based on the appearance of chemical materials. In addition, concepts corresponding to the actual names of chemical ingredients or substances which are either used for composing an HME recipe or constitute the outcome of a chemical process are also considered important. Indicative examples include concepts such as "nitric acid" or "styphnic acid". Moreover, concepts related to the texture of substances, e.g. transparent, translucent or opaque, are also considered relevant to the HME domain. Finally, the presence of an improvised explosive device or mechanism (consisting of many different handmade combined parts) is also considered as a plausible entity in HME multimedia content. For instance, parcel bombs, pressure cooker bombs or mobile phone bombs are possible objects that characterise the HME multimedia content and therefore should be detected so as to also enable the retrieval of the corresponding images/video shots.

B. HME-related concepts observed in empirical study

Another set of HME-related concepts emerged from an empirical study conducted by domain experts through manual examination and observation of HME multimedia content found on YouTube and relevant Web resources. In particular, a set of 13 HME YouTube videos and the images extracted from 48 HME Web resources, including Web pages, blogs, forums, and posts on social networks, all in the English language and all manually discovered, was thoroughly studied in a recursive manner based on the well-established empirical cycle methodology [39], so as to identify the meaningful HME recipe information they convey and the way this information is visually presented.

Our study showed that the vast majority of the multimedia content embedded in English HME Web resources could be characterised by the following. First of all, typically, there are people participating in the process of performing an HME recipe. The main concepts identified within this category are the ones depicting the number of people, the displayed human body-parts (e.g. arms, hands, fingers etc.), and rarely (in the case of the English multimedia content) people’s faces; even when people show their faces, they usually conceal their identity by wearing facial equipment (e.g. masks or sunglasses).

Moreover, the multimedia content comprises shots of the equipment used for composing an HME. Specifically, the concepts representing the equipment found are related to chemical glassware and utensils (e.g. glass vessels, measuring cups, test tubes etc.), appliances for heating (e.g. stove, slow cooker, etc.), pulverising (e.g. mortar and pestle), filtering (e.g. filters, funnels, etc.), and measuring (e.g. precision scales, syringes, droppers, etc.) the HME materials, as well as pieces of furniture (e.g. tables) serving as auxiliary equipment for performing recipes.

Furthermore, other useful concepts emanate from the location where the HME is being prepared or tested. Both of the aforementioned processes take place either in an indoor environment (e.g. room, kitchen, or laboratory) or outdoors (e.g. yard, garden, or woods). Another useful concept category that emerged from our study relates to the explosions shown on images or video frames that indicate the testing of the synthesised HME. These explosions may be accompanied by fire flames, smoke (coloured or not), and dust.

Our study also established that concepts related to specific visual elements present in images or video frames are also capable of characterising the HME multimedia content. Specifically, it found that overlaid text is commonly used in such video content either for communicating the HME recipe procedure or for introducing a viewer to the main content of a video chapter. It is also worth mentioning that many of the equipment used for composing HMEs contain readable textual information (a concept referred to as “scene text”). Moreover, the process of performing an HME recipe is usually depicted in video frames with a common static background and constitutes another important concept present in HME multimedia content. Finally, sketches containing graphical illustrations of HME devices, equipment, or ingredients may also occur on HME images, and hence they constitute another useful HME concept.

It is readily understood that several of the aforementioned HME-related concepts characterising the HME domain, have also a more generic application in many other different broader domains, from social interaction to cooking and travelling. This provides us with the opportunity of exploiting (annotated) multimedia datasets constructed in other domains. In particular, the TREC Video Retrieval Evaluation (<http://trecvid.nist.gov> – TRECVID) [40], an international benchmarking activity that supports research in concept-based video information retrieval, provides such large annotated test collections, along with uniform evaluation procedures, and a forum for organisations and research groups interested in comparing their results. Hence, we also consider HME-related concepts that are not only applicable to the HME domain,

TABLE I. SELECTED TRECVID CONCEPTS





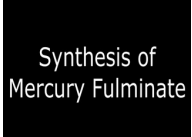

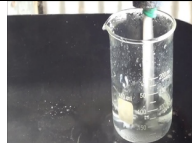




Concept	Description	Example
<i>Explosion or Fire</i>	Shots of an explosion or a fire	
<i>Attached Body Parts</i>	Shots of a close-up view of body parts attached to a live person. The head should not be visible; the body parts should be the primary focus	
<i>Indoor</i>	Shots depicting any interior scenery inside a room	
<i>Outdoor</i>	Shots of outdoor locations	
<i>Overlaid Text</i>	Shots of text that appears in the foreground, but the frame consists entirely of a computer graphics designed still image, and the text in question is part of the design	
<i>Scene Text</i>	Shots of text that is on a physical or virtual object such as a banner, button, or building	

TABLE II. SELECTED REMAINING HME-RELATED CONCEPTS

Concept	Description	Example
<i>Glassware</i>	Shots of glass vessels, e.g. beaker, flask, bowl, jar, test tube, vial, graduated cylinder, measuring cup, etc.	
<i>Explosion</i>	Shots depicting an explosion (with or without fire)	
<i>Liquid</i>	Shots of matter in fluid or in a semi-crystallised state either as a glassware’s content or as a flowing liquid	
<i>Smoke</i>	Shots depicting smoke of all colours (e.g. result of an explosion, a smoke bomb, burning material, etc.)	
<i>Sketch</i>	Shots of a drawing or painting depicting explosions, ingredients, devices, components, or other equipment related to HMEs	

but can also be encountered in the pool of the available TRECVID concepts. This allows us to either train concept detectors (or take advantage of concept detectors already trained) on the large volumes of TRECVID multimedia data, without having to create annotated datasets from scratch. Here, we focus on TRECVID concepts depicting locations (e.g. indoors, outdoors), people (e.g. human body parts, faces) and video elements (e.g. overlaid text, scene text, static background), since these are the categories we found relevant also to the HME domain.

C. Selected concepts

The previous sections introduced several concepts related to the HME domain that emerged either by the experts' contribution or by our empirical study. The concept selection for our evaluation experiments is based on the identification of the most representative concepts with respect to the available multimedia content. To this end, we have settled on a set of 11 concepts representing HME material, equipment, location, people, and sketches. The concepts belong either to the TRECVID generic concepts which can be applied to the HME domain, or to a list of specialised HME-oriented concepts. Table I presents the TRECVID concepts, whereas Table II describes the rest of the HME-related concepts.

V. EVALUATION

To assess the effectiveness of the proposed multimedia concept detection framework, a series of experiments are performed based on a dataset of HME-related multimedia content (images and videos). This section describes the experimental set-up of the evaluation procedure.

A. Datasets

The dataset used in our experiments consists of YouTube videos and also images embedded in Web resources (such as Web pages, blogs, and forums) containing HME recipes, with particular focus on recipes based on readily available materials. These HME videos and Web resources were manually or automatically discovered (using focussed crawling and search engine querying) and were manually assessed by domain experts as relevant to the HME domain.

In total, the dataset consists of 2900 items, 50 videos (decomposed into 2616 shots by the video segmentation process) and 284 images extracted from 726 Web resources. This set of images corresponds to the set generated after filtering out all the low-dimension images, (i.e. images not likely to contain HME-related visual information, mainly used for stylistic purposes). It should be noted that all the video shots are inherently considered as relevant to the HME domain (since they are extracted from relevant video resources), whereas the images may or may not be relevant to the HME domain.

B. Multimedia annotation

The dataset was manually annotated with respect to the 11 selected concepts. For the annotation process, a web-based annotation tool that presents the images/video frames of the dataset on the screen was employed. The annotator associated each image/video frame with the relevant concepts appearing beneath each item in a list of checkboxes. Given that most performance metrics (see next section) require binary relevance judgements, the

multimedia annotation was performed on the basis of two-dimensional space (i.e. each image/video frame is annotated either as relevant or as non-relevant to a concept). It should be noted that there is no limitation on the number of concepts that may appear on a single image, i.e. each image can be annotated as relevant to none or to more, even to all, concepts.

Table III lists the image/video frames annotation distribution per selected concept (i.e. number of relevant and non-relevant samples per concept). The concepts can be distinguished in three classes based on the number of relevant samples which they appear in: (1) highly frequent concepts (appearing in 20-30% of the dataset samples), including the concepts *Overlaid Text*, *Attached Body Parts*, *Liquid* and *Glassware*, (2) frequent concepts (appearing in 4-8% of the dataset samples), including the concepts *Indoor*, *Outdoor*, *Explosion or Fire*, *Scene Text*, and *Smoke*, and (3) rarely appearing concepts (appearing in less than 2% of the dataset samples), including the concepts *Explosion* and *Sketch*.

TABLE III. IMAGE DISTRIBUTION PER SELECTED CONCEPT; THE FIRST SIX CORRESPOND TO TRECVID CONCEPTS

Concept	Relevant	Non-relevant	Total
<i>Indoor</i>	124 (4.3%)	2776	2900
<i>Outdoor</i>	119 (4.1%)	2781	2900
<i>Explosion or Fire</i>	187 (6.4%)	2713	2900
<i>Overlaid Text</i>	715 (24.7%)	2185	2900
<i>Scene Text</i>	219 (7.6%)	2681	2900
<i>Attached Body Parts</i>	829 (28.6%)	2071	2900
<i>Explosion</i>	47 (1.6%)	2853	2900
<i>Smoke</i>	135 (4.7%)	2765	2900
<i>Liquid</i>	616 (21.2%)	2284	2900
<i>Glassware</i>	928 (32.0%)	1972	2900
<i>Sketch</i>	41 (1.4%)	2859	2900

C. Performance metrics

To evaluate the performance of the HME multimedia concept detection framework, we employ a set of reliable and widely used metrics. *Precision* (i.e. the proportion of the retrieved images/video frames relevant to each concept), *Recall* (i.e. the proportion of the relevant images/video frames retrieved per concept), and the *F-measure* (i.e. their harmonic mean) are employed. In addition, rank-based performance metrics are considered, including *precision at n* retrieved results ($P@n$), *R-precision* that calculates the precision at the top R images/video frames retrieved (where R is the number of relevant samples), *Average Precision* (AP), and the *Normalised Discounted Cumulative Gain* (NDCG) [41], a user-oriented metric that evaluates the usefulness, or gain, of the ranked result list.

D. Experimental set-up

Two sets of experiments are performed based on the concepts under consideration: (1) experiments on the six TRECVID concepts, and (2) experiments on the remaining five HME-related concepts. In the former case, the concept detectors are trained on the development dataset of the TRECVID 2014 Semantic Indexing Task consisting of 800 hours of video decomposed into 549434 shots and

annotated for several concepts, while our multimedia dataset (i.e. 284 images and 2616 shots) is used solely for testing purposes. The large size and heterogeneity of the dataset used for training the classifiers significantly reduces the potential for overfitting. On the other hand, for the remaining HME-related concepts, there is a need both to train the respective models and also to evaluate their effectiveness. Hence, the multimedia dataset is used both for training and testing (i.e. a part of the dataset is used for training and the remaining samples are used for testing). To this end, the initial video shot dataset is randomly divided into three distinct sets on a video basis (i.e. each set consists of all the video frames belonging to the videos included in it). The randomisation process is performed in such a way so as to guarantee that each set includes approximately one third of the initial video shots. Then, each set is enriched with one third of the samples present in the image set (the samples are randomly selected). To increase the generalisability of our findings, three experiments are performed based on a 3-fold cross validation where each of the three sets is used in turn for testing, while the other two are used for training. Table IV provides information concerning the datasets used for each of the three experiments both for training and testing. This random separation has resulted in preserving on average a relationship close to 2:1 among the number of relevant images in the training set against the number of relevant images in the testing set for every concept under consideration (see Table V and Table VI).

TABLE IV. NUMBER OF SAMPLES PER EXPERIMENT

	Train	Test	All
Experiment 1	1913	987	2900
Experiment 2	1946	954	2900
Experiment 3	1937	963	2900

Next, the evaluation results are presented.

VI. RESULTS

This section provides the evaluation results of the experiments performed for assessing the effectiveness of the framework for detecting the 11 HME-related concepts in multimedia content found on HME Web resources.

A. HME-related TRECVID concepts

Table V presents the results for the HME-related

TRECVID concepts. For all TRECVID concepts, precision improves significantly over the prior in the collection (reflected by the number of Relevant/Total), and, hence, the classifiers built demonstrate better performance in comparison to random classifiers. In most cases, precision and recall demonstrate the typical trade-off. In case the end user is interested in performing a precision-oriented search, *Explosion or Fire*, *Overlaid Text* and *Attached Body Parts* show particularly high performance over the entire set of retrieval results, while all concepts, apart from *Indoor*, also indicate very high precision at the top five and ten retrieval results. In addition, R-precision, AP and NDCG indicate the high effectiveness of our system on retrieving and ranking images/video frames relevant to *Overlaid Text*, while *Outdoor* and *Attached Body Parts* also demonstrate satisfactory ranking results. Figure 2 presents the top five images/video frames retrieved by our framework for the concepts *Overlaid Text* and *Attached Body Parts*.

In terms of recall, all concepts apart from *Explosion or Fire*, demonstrate satisfactory results. This could be due to the broad definition of this concept in TRECVID that probably results in being associated with particularly diverse visual content that may not be easily detected by a single concept detector. In terms of both precision and recall, the F-measure indicates that *Overlaid Text*, *Attached Body Parts* and *Scene Text* constitute the concepts that can be more accurately identified by our multimedia concept detection framework.

The considerable effectiveness for the concept *Overlaid Text* across all metrics indicates that it could be easily exploited for detecting images/video frames containing potentially useful text that could be extracted using OCR techniques. This would allow us to consider such text as an additional source of evidence both for assessing the relevance of the content of such resources, as well as for acquiring information not readily identifiable through the visual content alone, such as the names of the particular ingredients being used for the HME recipe. Moreover, the other well-performing concepts, such as *Attached Body Parts*, could be used in conjunction with other more specific HME-related concepts (see next section) for identifying multimedia content relevant to the HME domain.

Finally, Table V presents the average for all metrics across all TRECVID concepts. In general, these results

TABLE V. EFFECTIVENESS OF THE CONCEPT DETECTION FRAMEWORK FOR THE HME-RELATED TRECVID CONCEPTS

	Concepts						Average
	<i>Indoor</i>	<i>Outdoor</i>	<i>Explosion or Fire</i>	<i>Overlaid Text</i>	<i>Scene Text</i>	<i>Attached Body Parts</i>	
Relevant	124	119	187	715	219	829	365.50
Relevant/Total	0.04	0.04	0.06	0.25	0.08	0.29	0.13
Retrieved	1041	863	9	1167	208	1099	731.17
Precision	0.09	0.11	0.89	0.53	0.32	0.44	0.40
Recall	0.79	0.80	0.04	0.87	0.31	0.58	0.52
F-measure	0.17	0.19	0.08	0.66	0.31	0.50	0.32
P@5	0.40	1.00	1.00	1.00	1.00	0.80	0.87
P@10	0.40	1.00	0.80	1.00	0.80	0.90	0.82
R-precision	0.26	0.50	0.04	0.75	0.31	0.47	0.39
AP	0.21	0.51	0.04	0.78	0.17	0.34	0.34
NDCG	0.60	0.78	0.12	0.88	0.37	0.59	0.56

TABLE VI. EFFECTIVENESS OF THE CONCEPT DETECTION FRAMEWORK FOR THE REMAINING HME-RELATED CONCEPTS (AVERAGE OVER THREE EXPERIMENTS)

	Concepts					Average
	Explosion	Smoke	Liquid	Glassware	Sketch	
Training Set						
Relevant	31.33	90.33	410.67	620.67	29.33	236.47
Testing Set						
Relevant	15.67	44.67	205.33	307.33	11.67	116.93
Relevant/Total	0.02	0.05	0.21	0.32	0.01	0.12
Retrieved	2.67	11.67	79	265.33	10.33	73.80
Precision	0.56	0.43	0.52	0.64	0.78	0.59
Recall	0.13	0.10	0.20	0.55	0.69	0.33
F-measure	0.18	0.16	0.29	0.59	0.73	0.39
P@5	0.40	1.00	1.00	1.00	1.00	0.80
P@10	0.40	1.00	0.80	1.00	0.80	0.90
R-precision	0.13	0.10	0.20	0.55	0.69	0.33
AP	0.11	0.05	0.13	0.38	0.59	0.25
NDCG	0.20	0.16	0.28	0.58	0.73	0.39

indicate that our framework is capable of effectively detecting the aforementioned TRECVID concepts on HME visual content and returning decent results when it comes to ranked searches.

B. Remaining HME-related concepts

Regarding the remaining five HME concepts, Table VI presents the average evaluation results of the three experiments performed based on the 3-fold cross validation, so as to reach more reliable conclusions. Similarly to before, precision improves for all five concepts compared to the prior. In terms of precision, all

five concepts demonstrate satisfactory performance (precision fluctuates between 0.43 and 0.78 among all concepts). On the contrary, in terms of recall, the effectiveness of most concepts (with the exception of *Glassware* and *Sketch*) is not particularly high. Also, although most concepts demonstrate the expected trade-off among precision and recall, *Glassware* and *Sketch* show both high precision and recall. This indicates that our multimedia concept detection framework is capable of correctly identifying the respective objects on the vast majority of the cases. This is further confirmed when observing the effectiveness for these concepts in terms of

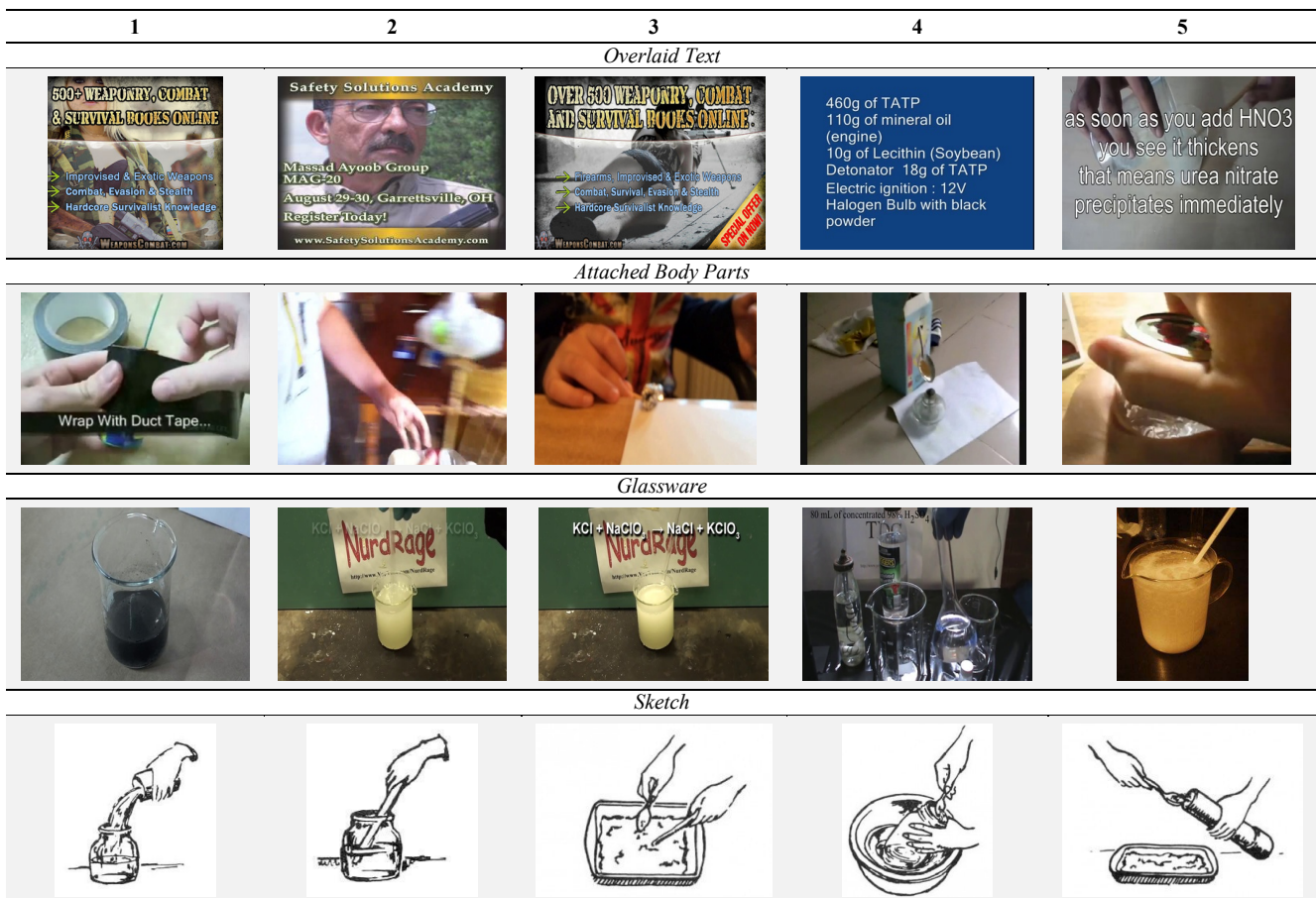


Figure 2: Top-five results retrieved for concepts *Overlaid Text*, *Attached Body Parts*, *Glassware*, and *Sketch*.

the remaining metrics (i.e. P@5, P@10, R-precision, AP, and NDCG). Figure 2 presents the top five samples returned by our framework for these two concepts in Experiment 1.

Therefore, the presence of *Glassware* in conjunction with the other high precision concepts, such as *Smoke*, and *Liquid* as well as *Attached Body Parts* could be used for identifying multimedia content and events relevant to the HME domain. Furthermore, it should be noted that for the less frequent concepts (i.e. *Explosion* and *Sketch*), further experiments with larger numbers of relevant samples are needed for reaching more reliable conclusions. Finally, Table VI presents the average results for all metrics across all concepts among all three experiments. Similarly to before, these results indicate the satisfactory effectiveness of our framework for detecting HME-related concepts on multimedia content and retrieving such content on a concept basis.

VII. CONCLUSIONS

This work presented an HME multimedia concept detection framework, built on one hand for identifying HME-related visual content on the Web, and on the other hand for supporting end-user queries on such HME-related. The framework is based on state-of-the-art approaches in the fields of video segmentation, feature extraction and classification. Its effectiveness has been tested on multimedia content (both videos and images) embedded in HME Web resources. The concepts under consideration belong either to the generic TRECVID concepts that are also applicable to the HME domain or to other specialised HME-related concepts. We demonstrated that our framework is capable of detecting most of these concepts in visual content with satisfactory precision. We also showed that our framework is able to return rank-based results for supporting end-user searches with decent performance in most cases. To sum up, the HME multimedia concept detection framework has indicated the significant adaptability of the adopted approaches to the HME domain. Future work includes the further evaluation of the concept detection framework on larger more heterogeneous collections and with respect to additional HME-related concepts, as well as the examination of event detection approaches based on the detected concepts.

ACKNOWLEDGMENT

This work was supported by the HOMER (312388), MULTISENSOR (610411) and ForgetIT (600826) FP7 projects partially funded by the European Commission.

REFERENCES

- [1] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, 2004, pp. 91–110.
- [2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision & Image Understanding*, vol. 110(3), 2008, pp. 346–359.
- [3] K. Van de Sande, T. Gevers, and C. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 32(9), 2010, pp. 1582–1596.
- [4] F. Markatopoulou, N. Pittaras, O. Papadopoulou, V. Mezaris, and I. Patras, "A Study on the Use of a Binary Local Descriptor and Color Extensions of Local Descriptors for Video Concept Detection," *Proc. 21st International Conference on Multimedia Modeling (MMM 15)*, 2015.
- [5] F. Markatopoulou, V. Mezaris, N. Pittaras and I. Patras, "Local Features and a Two-Layer Stacking Architecture for Semantic Concept Detection in Video," *IEEE Transactions on Emerging Topics in Computing*, vol. 3 (2), 2015, pp. 193-204.
- [6] C. Olston and M. Najork. *Web Crawling. Foundations and Trends in Information Retrieval*, vol. 4(3), 2010, pp. 175-246
- [7] C. Snoek and M. Worring, "Concept-Based Video Retrieval," *Foundations and Trends in Information Retrieval*, vol. 2, 2008, pp. 215-322.
- [8] W. Tan, S. Teng, and W. Zhang, "Research on Video Segmentation via Active Learning," *Proc. Fourth International Conference on Image and Graphics (ICIG 07)*, IEEE Computer Society, 2007, pp. 395-400.
- [9] S. Liu, M. Zhu, and Q. Zheng, "Video Shot Boundary Detection with Local Feature Post Refinement," *Proc. 9th International Conference on Signal Processing (ICSP 08)*, 2008, pp. 1548–1551.
- [10] X. Ling, O. Yuanxin, L. Huan, and X. Zhang, "A Method for Fast Shot Boundary Detection based on SVM," *Proc. Congress on Image and Signal Processing (CISP 08)*, vol. 2, 2008, pp. 445–449.
- [11] E. Apostolidis and V. Mezaris, "Fast Shot Segmentation Combining Global and Local Visual Descriptors", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [12] E. Tsamoura, V. Mezaris, and I. Kompatsiaris, "Gradual Transition Detection Using Color Coherence and Other Criteria in a Video Shot Meta-Segmentation Framework," *Proc. 15th IEEE International Conference on Image Processing (ICIP 08)*, 2008, pp. 45–48.
- [13] M. Yeung, B. Yeo, and B. Liu, "Segmentation of Video by Clustering and Graph Analysis," *Computer Vision & Image Understanding*, vol. 71(1), 1998, pp. 94–109.
- [14] Z. Rasheed and M. Shah, "Scene Detection in Hollywood Movies and TV Shows," *Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 03)*, vol. 2, 2003, pp. 343–348.
- [15] J. Liao and B. Zhang, "A Robust Clustering Algorithm for Video Shots Using Haar Wavelet Transformation," *Proc. SIGMOD 2007 Workshop on Innovative Database Research (IDAR 07)*, 2007.
- [16] Y. Zhai and M. Shah, "Video Scene Segmentation Using Markov Chain Monte Carlo," *IEEE Transactions on Multimedia*, vol. 8(4), 2006, pp. 686-697.
- [17] A. Bosch, A. Zisserman, and X. Munoz, "Scene Classification Using a Hybrid Generative/ Discriminative Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30(4), 2008, pp. 712–727.
- [18] J. Van de Weijer, T. Gevers, and A. Bagdanov, "Boosting Color Saliency in Image Feature Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28(1), 2006, pp. 150–156.
- [19] R. Tao, "Visual Concept Detection and Real Time Object Detection," *CoRR*, abs/1104.0582, 2011.
- [20] J. Sivic and A. Zisserman, "Efficient Visual Search for Objects in Videos," *Proc. IEEE*, vol. 96, 2008, pp. 548–566.
- [21] G. Qiu, "Indexing Chromatic and Achromatic Patterns for Content-Based Colour Image Retrieval," *Pattern Recognition*, vol. 35, 2002, pp. 1675-1686.
- [22] K. Van de Sande, C. Snoek, and A. Smeulders, "Fisher and Vlad with Fair," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [23] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The Devil is in the Details: an Evaluation of Recent Feature Encoding Methods," *Proc. British Machine Vision Conference*, 2011, pp. 76.1-76.12.
- [24] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," *Proc. 11th European Conference on Computer Vision: Part IV*, 2010, pp. 143-156.

- [25] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating Local Image Descriptors into Compact Codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34(9), 2012, pp. 1704-1716.
- [26] M. Guironnet, D. Pellerin, and M. Rombaut, "Video Classification Based on Low-Level Feature Fusion Model," *Proc. 13th European Signal Processing Conference*, 2005.
- [27] T. Leon, P. Zuccarelli, G. Ayala, E. de Ves, and J. Domingo, "Applying Logistic Regression to Relevance Feedback in Image Retrieval Systems," *Pattern Recognition*, vol. 40, 2007, pp. 2621.
- [28] D. Cireşfan, U. Meier, J. Masci, L. Gambardella, and J. Schmidhuber, "Flexible, High Performance Convolutional Neural Networks for Image Classification," *Proc. 22nd Intl. Joint Conf. on Artificial Intelligence*, 2011, pp. 1237-1242
- [29] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional Networks and Applications in Vision," *Proc. IEEE International Symposium on Circuits and Systems*, 2010, pp. 253-226.
- [30] M. Montagnuolo and A. Messina, "Automatic Genre Classification of TV Programmes Using Gaussian Mixture Models and Neural Networks," *Proc. 18th International Conference on Database and Expert Systems Applications*, 2007, pp. 99-103.
- [31] C. Lu, M. Drew, and J. Au, "Classification of Summarized Videos Using Hidden Markov Models on Compressed Chromaticity Signatures," *Proc. 9th ACM International Conference on Multimedia*, 2001, pp. 479-482.
- [32] X. Gibert, H. Li, and D. Doermann, "Sports Video Classification Using HMMS," *Proc. 2003 International Conference on Multimedia and Expo*, 2003, pp. 345-348.
- [33] N. Rota and M. Thonnat, "Video Sequence Interpretation for Visual Surveillance," *Proc. Third IEEE International Workshop on Visual Surveillance (VS 00)*, 2000, p. 59.
- [34] M. Bhargava, C. Chen, M. Ryoo, and J. Aggarwal, "Detection of Abandoned Objects in Crowded Environments," *Proc. 2007 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 07)*, 2007, pp. 271-276.
- [35] W. Chmiel, J. Kwiecień, and Z. Mikrut, "Realization of Scenarios for Video Surveillance," *Image Processing & Communications*, vol. 17(4), 2013, pp. 231-240.
- [36] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti, "Attribute-Based People Search: Lessons Learnt from a Practical Surveillance System," *Proc. International Conference on Multimedia Retrieval (ICMR 14)*, 2014, pp. 153-160.
- [37] H. Chen (Ed.). "Dark Web: Exploring and Data Mining the Dark Side of the Web", Springer, 2012.
- [38] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf", 2011 IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2564-2571.
- [39] A. D. de Groot, "Methodology. Foundations of inference and research in the behavioural sciences". Mouton, The Hague, The Netherlands, 1969.
- [40] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID. Proc. 8th ACM International Workshop on Multimedia Information Retrieval (MIR '06)", 2006, pp. 321-330.
- [41] K. Järvelin and J. Kekäläinen. "Cumulated gain-based evaluation of IR techniques." *ACM Transactions on Information Systems (TOIS)* vol. 20(4), 2002, pp. 422-446.