Automatic Segmentation, Aggregation and Indexing of Multimodal News Information from Television and the Internet

Maurizio Montagnuolo[1], Alberto Messina[1,2], Roberto Borgotallo[1]
[1]RAI Radiotelevisione Italiana
Centre for Research and Technological Innovation
C.so Giamone 68, I-10135 Torino (Italy)
{maurizio.montagnuolo, a.messina, r.borgotallo}@rai.it

[2]Università degli Studi di Torino
Department of Computer Science
C.so Svizzera 185, I-10149 Torino (Italy)
messina@di.unito.it

Journal of Digital Information Management

ABSTRACT: *The global diffusion of the Internet has enabled the distribution of informative content through dynamic media such as RSS feeds and video blogs. At the same time, the decreasing cost of electronic devices has increased the pervasive availability of the same informative content in the form of digital audiovisual data. This article presents a system for the large-scale unsupervised acquisition, segmentation and indexing of TV newscasts. In particular, it discusses the principles and performance of the parts of the system dedicated to the detection and segmentation of programmes from the acquired stream. In addition to the core technology, we also introduce and discuss a novel method for assessing the results of story boundaries segmentation algorithms, based on a user-validated measurement. Due to the heterogeneity of current news distribution channels, a further innovative aspect of this article is the description of a framework for multimodal information aggregation. The core of this framework is a cross-modal clustering process for which a novel, asymmetric similarity measure is provided. The implemented prototype uses online news articles and TV news programmes as information sources, and provides a multimodal service integrating both contributions. Experimental evaluation of the system proves the effectiveness of the method in the studied case.*

Categories and Subject Descriptors
H. 4.0[Information Systems Applications]: General; **H 3.1[Information Storage and Retrieval]** Content Analysis and Indexing; **I 5.3[Pattern Recognition]** Clustering

General Terms: Web Indexing, Internet, TV News Indexing

Keywords: Content segmentation, Multimodal aggregation, Multimedia annotation, Semantic enrichment

Received: 11 May 2010: Revised 17 June 2010; Accepted 27 June 2010

1. Introduction and Related Work

Modern information society is characterised by a rapid evolution of information consumption models, and by an overwhelming amount of multimedia data being produced every day, and delivered through several multimodal information channels, among which digital interactive television is still a major one. It is now a commonplace that interactive television, in its different and variegated interpretations and embodiments (e.g., IPTV, Web TV) is seen as the next frontier of media production businesses. However, exploiting the full potential of putting users in the loop is still jeopardised by the limited capability of traditional broadcasters of turning their established workflows into something adapted to the new scope. In such a highly dynamic context, the convergence of Internet technology and digital television represents a principal driver towards the *automatisation* of all the production processes to fulfil the requirements of the new emerging application domains.

As a first step in such automatisation process, automatic programme segmentation concerns the ability of automatically detect semantically coherent parts of television programmes. This step is a key enabler for all interactive applications that include informative content management, since the ability of correctly and efficiently segmenting news content is the basis on which several other application can be developed, e.g. recommendation systems, users preferences collection profiling, personalised home TV applications. Some reference works in this area are those presented in [8, 10, 15]. In particular, to solve the *news story* segmentation, the common base of the approaches is constituted by the use of a combination of visual, audio and speech features. The TRECVID initiative had news segmentation among its tasks in 2003 and 2004. The works in [7, 12] present the various approaches identified and developed by the TRECVID participants in those two series. These approaches included either video and audio channels analysis or, in addition, speech-to-text automatic transcripts. The baseline features employed in several cases are visual similarity between shots within a time window and the temporal distance between shots, e.g. [9]. Other heuristics like similarity of faces appearing in the shots and the detection of the repeated appearance of anchorperson shots can add a supplemental layer of information to improve the overall accuracy [15, 19, 22]. The audio channel contribution can be employed to detect pauses, potential boundaries for topic changes [9, 16, 19], or to detect changes in audio classification patterns (e.g., music to speech changes [9, 16]), or to detect speaker changes [16]. As a third information source, text from transcripts or automated speech recognition is very often used, either by searching similar word appearances in different shots or by detecting text similarities between the shots [15, 19].

A first critical aspect connected with these technologies is related to their effective and efficient implementation in real-life domains, where operating conditions and data characteristics may significantly vary from prototype environments. Another critical aspect is specifically connected with the segmentation algorithm, i.e. the rigorous and robust evaluation of its accuracy. Evaluation systems based on exact boundary matches may be

little reliable on expressing real effectiveness of a segmentation algorithm. In fact, they do not consider any intermediate relevance, and a news item is defined correctly retrieved only if it is exactly matching the reference values, even if there is only a few seconds time lag between detected and true segments. The TRECVID evaluation strategy extends this approach by adding a fixed uncertainty threshold of 5 seconds to each boundary. However, this symmetric window lack in taking into account the different sensitivity of users to starting and ending story boundaries, as well as the different objective weight that has to be assigned to extra or missing material.

A second critical aspect regards the methodologies used to aggregate and present the large amounts of information available from different sources and heterogeneous media formats. In fact, the global diffusion of the Internet and the evolution of Web technologies are enabling the creation and delivering of dynamic informative contents such as news, blogs or vodcasts. Those contents are usually distributed through RSS feeds, an XML-based format originally proposed by Netscape to provide simple, extensible and flexible content distribution [1]. Users can manage RSS feeds using a feed reader that periodically downloads the updated contents from the subscribed feeds, displays the items in each feed and provides links to the related resources. However, returning the unorganised list of all items included in the subscribed feeds ends in causing an information overload effect. One solution to address this problem consists in aggregating similar items sharing some common attributes. Pioneer works used lists of keywords to describe the feeds contents [11]. More recent approaches use clustering techniques and knowledge integration [6, 13]. The common background of such approaches is their capability of performing automated aggregation of data collected from a single, closed domain, i.e. the Internet. So far, however, only a few attempts have been made to investigate the possibility of merging information from heterogeneous information channels, such as television broadcasts and Web pages. As an example, the work in [21] uses both internal audiovisual features and different external information sources for event detection in team sports video. Similarly, the same task is performed in [20, 23], by making use of web-casting text and broadcast video information.

Further exploring this direction, this article describes an unsupervised framework for content-based TV news stories and online newspaper articles aggregation and retrieval. The rest of the article is organised as follows. Section 2 describes the architecture of our prototype system. Section 3 presents experimental evaluations. Final conclusions and future work are illustrated in Section 4.

2. Prototype System Architecture

Figure 1 illustrates the high-level architecture of our prototype system. The system can be viewed as a processing machine with two input channels, i.e. digitised broadcast news streams (DTV) and Web RSS feeds (RSSF), and one output channel, i.e. the multimodal aggregation service (MMAS), which is automatically determined from the semantic aggregation of the input streams.

2.1 DTV Processing Chain

The DTV processing sub-system was designed and implemented to offer an efficient, scalable and robust software architecture for the acquisition and automatic annotation of broadcast news, including automatic news story segmentation and semantic analysis of spoken text with Named Entities Recognition and Classification (NERC). Firstly, the broadcast news streams are automatically detected and segmented into their elementary

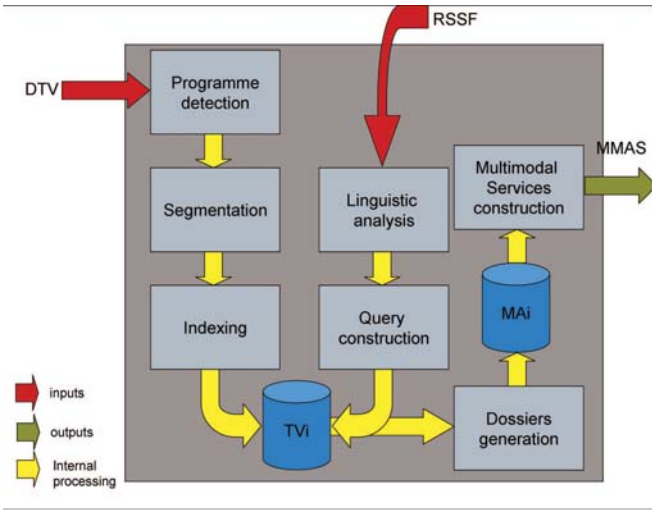

Figure 1. High-level prototype system architecture

news stories. The detection is performed using the automatic video clip detection technique described in Section 2.1.1. The segmentation process is done by exploiting aural and visual cues, with the help of a three-layered heuristic framework, as presented in Section 2.1.2. Once segmented, the audio track of each story is transcribed and indexed in the TVi catalogue (see Section 2.1.3), which serves as the permanent repository from which full text search and category search functionalities are delivered to the final users.

2.1.1 Detection of programmes from live streams

To achieve automatic segmentation of live streams into programmes we make use of an optimised video clip matching technique. Video elements (shots) indicating starting and ending of programmes are used as reference prototypes to be searched through the acquired video streams. The technique consists of a learning phase followed by a detection phase.

In the learning phase, for each macro-event of interest E, e.g. a programme starting or ending jingle, N_E instances $\{e_1, e_2,..., e_{N_E}\}$ are selected from daily television programme acquisitions. On each instance e_m an adaptive threshold shot detection algorithm based on displaced frame luminance difference (L-DFD) is performed. Adaptivity is based on the local statistics of L-DFD, so that content having higher L- DFD variance is processed against higher thresholds.

After the shot detection process, each clip instance e_m is split into N_s shots forming the set $S_m = \{S_1, S_2,..., S_{N_S}\}$, whose length array is $l_{Sm} = (l_1, l_2,..., l_{N_S})$, where l_n, $n = 1,..., N_s$ is the number of frames within the shot S_n. Let $F = (f_1, f_2,..., f_{N_F})$ be the set of low-level visual features extracted from each shot, which includes hue, saturation, and value colour descriptors, luminance, contrast and directionality texture descriptors [18], temporal activity motion descriptor. Each element of F is represented using a uniformly distributed B-bin histogram, where the last bin is used to count the pixels of the frame for which the measurement of feature returns an undetermined value (e.g., hue for grey pixels). Therefore, each macro-event E is represented by the set $A_E = (a_1, a_2,..., a_{N_E})$, which, in turn, is made of N_E arrays of feature vectors of size Ns, Each element a_{mn}, $n = 1…Ns$ of each array $a_m \in A_E$ is a real matrix of dimensionality $N_f \times B \times l_n$, where we recall that N_f denotes the total number of extracted low-level features (i.e., $N_f = 7$), B denotes the total number of bins used in each feature histogram (i.e., $B = 65$), and l_n denotes the length (in frames) of the shot S_n.

To improve the efficiency of the detection, a selection of the *most promising* shots among the Ns shots pertaining to the macro-event E is performed. The selection is based on the measurement of the Bhattacharya distance between the distribution of the feature vectors extracted from each of the *Ns* shots and a reference set of shots $R = \{r_1, r_2,..., r_{N_R}\}$ that represents the generic population of shots among which the macro event E has to be searched. During the shot selection, a feature selection is also performed, in order to pick up the features maximising the measured divergence. More formally, we select a subset of the set of detected shots *S* and a subset of the set of extracted features F, based on the value assumed by the following function:

$$B_{jk} = \frac{1}{2}\sum_{m=1}^{B} \ln \frac{\sigma^2_{jkm} + \sigma^2_{rkm}}{2\sqrt{\sigma^2_{jkm}\sigma^2_{rkm}}} + \frac{1}{4}\sum_{m=1}^{B} \frac{\left|\mu_{jkm} - \mu_{rkm}\right|}{\sigma^2_{jkm} + \sigma^2_{rkm}} \quad (1)$$

$$\mu_{jkm} = \frac{1}{l_j}\sum_{h=1}^{l_j} \frac{1}{N_E}\sum_{l=1}^{N_E} a_{lj}^{kmh} \quad (2)$$

$$\sigma_{jkm} = \frac{1}{l_j}\sum_{h=1}^{l_j} \sigma_{l_j}^{kmh} = \frac{1}{l_j}\sum_{h=1}^{l_j} \sqrt{\frac{1}{N_E}\sum_{l=1}^{N_E}\left(a_{lj}^{kmh} - \frac{1}{N_E}\sum_{l=1}^{N_E} a_{lj}^{kmh}\right)^2} \quad (3)$$

Under the assumption that the elements of the N_f extracted feature histograms follow independent Gaussian distributions, the elements Equation (1) is an approximation of the Bhattacharya distance [26] between the distributions of the features in the set of instances of the shots of the macro-event E and the distributions of the features in the general population of all shots. Indices *j;m* vary in $1…B$, while indices *k;h* vary in $1…N_f$ and $1…l_n$, respectively.

The value of the coefficient B_{jk} (see Equation 1) can be taken as a measurement of the distinctiveness of feature $f_k \in F$ for the shot $s_j \in S_m$. Following this assumption we can therefore select the top-*K* shots among the available *Ns* shots w.r.t. each of the N_f available features. This selection can be represented arranging the elements B_{jk} in a matrix **B** and selecting the couples (s_{a_i}, f_{b_j}), $s_{a_i} \in S_m$, $f_{b_j} \in F$, corresponding to the elements of **B** reaching the top-*K* values. This approach enhances the precision of the detection phase, since it selects the combinations shot/feature that maximise divergence of the statistical distributions of reference shots w.r.t. a generic population. However, this is not yet enough to ensure the optimisation of recall, since selected feature distributions may have significant standard deviations around their mean values. To limit this drawback, we select the best *combination* of features among the total possible $\sum_{i=1}^{N_f}\binom{N_f}{i}$ combinations, in order to exploit diversity effects of different features. The combination is chosen as the one maximising the harmonic mean of the divergence measurement given by the individual features corresponding to the B_{ij} coefficients.

In the detection phase, averaged feature vectors $\langle a^{ij} \rangle = \frac{1}{l_n}\sum_{k=1}^{l_n} \frac{1}{N_E}\sum_{l=1}^{N_E} a_{\ln}^{ijk}$ are used as fixed references, with respect to which acquired shots are compared using a distance measurement based on histogram intersection. Let *I* be an acquired shot and $s_J \in S$, $J \in 1...N_s$ a reference shot for the macro-event *E*, and $\{a_I^{ij}, i=1...N_{sel}\}$ and $\{a_{s_J}^{ij}, i=1...N_{sel}\}$ their respective sets of averaged feature vectors, where N_{sel} is the number of selected features for the reference shot s_J. The distance $D(I, s_j)$between l and s_J is defined as follows:

$$D(I, s_J) = \left[\prod_{i=1}^{N_{sel}} d_i(I, s_J)\right]^{\frac{1}{N_{sel}}} \quad (4)$$

$$d_i(I, s_J) = \sum_{j=1}^{B} \min\left(a_I^{ij}, a_{s_J}^{ij}\right) \quad (5)$$

The acquired shot I is classified as an instance of the reference shot s_J if:

$$D(I, s_J) \le \alpha\,\sigma_J^{ij} = \frac{\alpha}{l_J}\sum_{k=1}^{l_J} \sigma_J^{ijk} \quad (6)$$

where l_J is the length of the shot and $\alpha \in [0,1]$ is a parameter governing the recall of the detection.

2.1.2 Segmentation of TV Newscasts

After having acquired and detected programme boundaries following the approach described in Section 2.1.1, newscast programmes are imported into the system and segmented into their logical units, i.e. news stories, which are the elementary items stored in the MAi index (see Figure 1). Segmentation of newscasts programmes into news stories is done exploiting aural and visual cues with the help of a three-layered heuristic framework. The used heuristics are based on the observation of the stylistic language of a set of 80 TV newscasts, taken in a controlled period of time from daily schedules of the 7 major national broadcast channels.

The basic heuristic (H1), also adopted in literature by e.g. [8], consists in considering boundaries of shots containing the anchorman as equivalent to news stories boundaries.

In order to detect the anchorman shots we use a second heuristic (H2), consisting in observing that the most frequent speaker is the anchorman and that (s)he speaks many times during the programme, and for periods of time distributed all along the programme timeline. This observation allows selecting the speaker who most likely is the anchorman, provided that a speaker clustering process labels all the speakers present in the programme and associates them to temporal segments of the content. However, the application of the first two heuristics is not yet enough to discern situations where the anchorman introduces several brief stories in sequence, without interruptions filled with external contributions. To overcome this limitation we use the third heuristic (H3), i.e. knowing that in the great majority of observed cases the introduction of a new brief story is accompanied by a camera shot change (e.g., from a close up shot to a wider one).

Thus, to optimise the accuracy of segmentation, we perform a shot clustering process based on both audio segmentation [14] and video features [24]. This allows us to detect and classify shot clusters as pertaining to studio shots containing the anchorman following the same frequency/extension heuristic used for detecting the candidate speaker (H2). This double clustering process enables a very simple and effective algorithm that selects video and audio clusters on the basis of their mutual coverage percentage.

More formally, let $C_A = \{a_1, a_2,..., a_{N_A}\}$ and $c_V = \{v_1, v_2,..., v_{N_V}\}$ be the set of speaker clusters and the set of video shot clusters for a detected newscast programme, respectively. Each element of C_A (and similarly of C_V) is represented as $a_i = \{\forall i, j : ext_{ij}^a = [ts_{ij}, te_{ij}]\}$, being ts_{ij}, te_{ij} respectively the starting and ending times of the j-th element of a_i w.r.t. the programme timeline. Be $\varepsilon(a_i)$ a function returning the extension of cluster a_i, defined as:

$$\varepsilon(a_i) = \max_j(te_{ij}) - \min_j(ts_{ij}) \quad (7)$$

We select $a_i \in C_A$ (and similarly $v_k \in C_V$) as a pivotal cluster if the following condition parameterised by the threshold τ_p holds:

$$\varepsilon(a_i) \cdot |a_i| > \tau_p \quad (8)$$

Be $C_A^p \subseteq C_A$ and $C_V^p \subseteq C_V$ the subsets of clusters containing only the pivotal elements selected through Equations (7) and (8). We further select among elements of C_V^p, the top-M scoring elements $v_1, v_2,..., v_M$ w.r.t. the temporal coverage with elements of C_A^p, defined as:

$$COV(v_l, a_i) = \frac{1}{|a_i||v_l|}\sum_{k=1}^{|a_i|}\sum_{m=1}^{|v_l|} length(e_{ik}^a \cap e_{lm}^v) \quad (9)$$

where $length(\cdot)$ returns the duration of its argument as per the intuitive definition. Let C_V^{p*} be the further pruned version of C_A^p, after the selection of the top-M elements. Then, story boundaries are identified as those in which an audio or a video cluster boundary occurs among the elements of $C_{AV} = C_A^p \cup C_V^{p*}$, with a threshold τ_o to avoid over-segmentation. Namely, elements $e^{a,v}$ of all members of C_{AV} are at first ordered w.r.t. their starting times ts, forming a vector $(e_1^{a,v}, e_2^{a,v},..., e_N^{a,v})$. Then, an element $e_k^{a,v} = [ts_k, te_k]$ is discarded if $ts_k, ts_{k-1} < \tau_o$. To sum up, the parameters of the segmentation algorithm are: τ_p, M and τ_o.

Figure 2 illustrates an example. The anchorman shots a and b are detected according to the heuristic $H2$ because both contain the same speaker A. As a shot boundary is detected between the shots a and b, the first two stories are segmented according to $H3$. The succeeding stories are then detected according to $H1$.

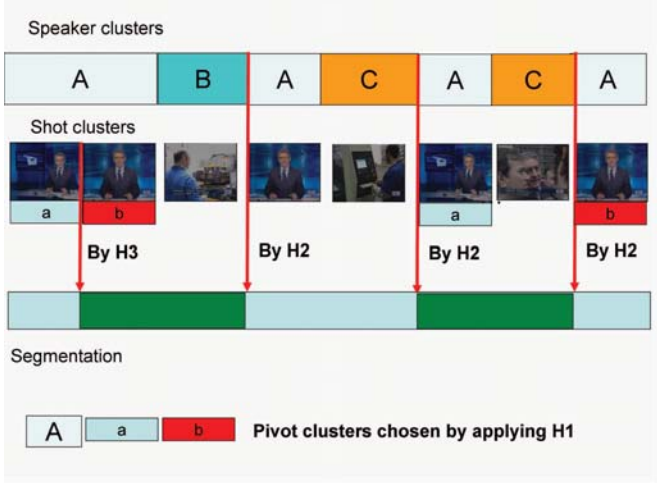

Figure 2. Illustration of news story segmentation

2.1.3 News Stories Indexing

Spoken content of the detected news stories is extracted using a speech to text engine based on [4], capable of translating both Italian and English. Subject classification of stories is done using a naive Bayesian classification model trained on a corpus made up of items of extracted text and annotated according to a taxonomy of 28 categories. The corpus counts 25,000 items, 4/5 of which were used for training and the remaining 1/5 for test. Overall subject classification accuracy is 0.82, while programme level accuracy, i.e. average classification accuracy calculated on items belonging to the same programme, is 0.88. Named entity recognition from transcribed text is also performed, using the technology described in [5].

2.2 RSSF Processing Chain

The RSSF stream is constituted by the RSS feeds of several major online newspapers and press agencies. An RSS feeds in constituted by a set of *items*, each described by a *title* (i.e., the headline of the corresponding online article), a *description* (i.e., a brief summary of the content of the corresponding online article), a *publication* timestamp (i.e., the date and time when the article was first published on the newspaper Web site) and a *link* to the corresponding full article. Additional metadata could be also included, such as a list of item's categories, or some comments, according to the RSS specifications [1]. Each RSS item can be expressed as a tuple $\ell = (uuid, pubDate, link, feed, title, phrases)$, where *uuid* is the universal identifier of the item, *feed* is the name of the source feed where the item was found and *phrases* is the set of sentences included in the item's description. All such tuples are stored as records in a PostgreSQL database.

On each RSS item a linguistic analysis based on POS tagging [17] is performed to identify the linguistic entities, e.g. verbs, nouns, adjectives, included in the title and the description phrases of the RSS item. The outputs of the linguistic analysis are employed by the query constructor to generate a set of representative query expressions, as detailed in Section 2.2.1. The generated queries are then submitted to the index structure of the TVi documents catalogue. For each item, the result of this search operation is a weighted set of newscasts stories of decreasing affinity to the target query.

On the results of such queries, a cross-modal clustering process is performed to aggregate items based on their semantic similarities, thus producing what we call the "*multimedia dossier*", i.e. a multimodal aggregation of TV newscast stories and newspapers articles sharing the same semantic content. Further analytical details on the clustering process are provided in Section 2.2.2.

For each multimedia dossier a unique identifier is provided in addition to the transcriptions of all the included news stories and the title and descriptions of all the included RSS items. This information is finally stored in the multimodal aggregation index (MAi), which is based on Lucene [25] and it serves as the permanent database from which the multimodal aggregation service is maintained and delivered to the final users. The system supports full-text queries and even field specific queries (e.g. publication/broadcast date, category, RSS provider, broadcast channel), allowing users to search across all the available data. To facilitate the results visualisation, the system provides a Web interface showing the ranked results. The details of the browsing interface are presented in Section 2.2.3

2.2.1 Linguistic Analysis and Query Building

In our system RSS feeds are used as the starting point for the aggregation process. Our system works with both Italian and

English language feeds, using an automatic translation service based on the OpenLogos tools [2]. The RSS items are analysed by TreeTagger [3], a language independent part-of-speech software that tokenises the input text (i.e., the items' titles and description phrases) and tags each word of the text with a descriptive label, according to a key set of grammar categories.

Let π be an RSS item described by the tuple ℓ. The linguistic analysis module first splits the text of the item's title and description phrases into *m* independent sentences. Then, it maps all such sentences into a vector of key/value pairs, $\pi((k,v)_{l1},...,(k,v)_{lf1},...,(k,v)_{lm},...,(k,v)_{lm})$, where the keys are the words in the sentences, the values are the corresponding grammar categories, *f* and *l* are the number of words in the first (i.e., the item's title) and last sentence (i.e. the last item's description phrase), respectively.

Based on this vector, the query construction process works in two steps. First, for each sub-vector $s_i \in \pi$, a basic query q_i is built selecting the words in s_i tagged as nouns. Then, a complex query *Q* is generated, joining all the basic queries as follows:

$$Q := \bigcup_{i=1}^{m} q_i^{2^{(m-i)}} \quad (10)$$

Equation (10) associates higher weights to queries derived from sentences occurring earlier, in order to emphasise the title and the initial description sentences. An example of the query construction process is shown in Figure 3.

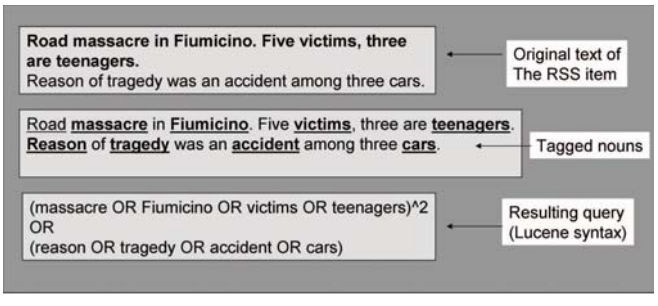

Figure 3. Example of query construction from linguistic analysis

2.2.2 Cross-Modal Clustering for RSS Items and News Stories Aggregation

The task of RSS items and news stories aggregation is addressed using a cross-modal clustering algorithm. The process is implemented in four steps:

1. RSSF-DTV affinity matrix building;
2. RSSF equivalence matrix derivation;
3. RSSF connectivity graph generation;
4. Multimedia dossiers construction and indexing.

In the first step, a *search results vector* is computed for each RSS item. Let $W_{eb} = \{\pi_i\}_{i=1}^{m}$ be the set of downloaded RSS items, and $T_v = \{N_j\}_{j=1}^{n}$ be the set of broadcasted news stories. For each RSS item $\pi_i \in W_{eb}$ the corresponding complex query Q_i is launched on the news stories speech transcriptions stored in the TVi index structure. The output of Q_i is stored in the search results vector $S_i = (S_{i1},..., S_{in})$, where S_{ij} is the Lucene score[1] of the query Q_i to the speech transcription of the news story N_j. We call the score S_{ij} as the *affinity* of the news story N_j to the RSS item π_i. All the search results vectors are then arranged in the rows of the *affinity matrix* $A = [s_1,..., s_i,..., s_m]^T \in [0,1]^{m \times n}$. The matrix *A* can be considered as a *space transformation* operator that projects the RSS items from the Web space, i.e. the text space of the complex queries, to the broadcast domain, i.e. the multimedia space of the TV news stories.

Once the affinity matrix has been constructed, we compute the similarity between RSS items exploiting their projection in the broadcast domain (step 2). The similarity between the couple of RSS items $(\pi_a,\pi_b) \in W_{eb}$ is evaluated using an asymmetric affinity function $s(\pi_a,\pi_b)$, as defined in the following equations:

$$S(\pi_a,\pi_b) = \cos(s_a,s_b)\frac{\|s_b\|}{\|s_a\|} \quad (11)$$

$$\text{iff } S(\pi_a,\pi_b) > \alpha \wedge S(\pi_a,\pi_b) > \alpha \text{ then } Eq(\pi_a,\pi_b) \quad (12)$$

$$\text{iff } S(\pi_a,\pi_b) > \alpha \wedge S(\pi_a,\pi_b) \leq \alpha \text{ then } Ent(\pi_a,\pi_b) \quad (13)$$

where s_a and s_b are the search results vectors obtained in the first step and $\cos(s_a, s_b)$ is the Cosine similarity between s_a and s_b. Intuitively, the function $S(\cdot)$ in Equation (11) measures *how much* the item π_a is explained by the item π_b, in the space of their search results vectors. Equation (12) defines the *semantic equivalence* relation between π_a and π_b, while Equation (13) defines the *semantic entailment* relation from π_a and π_b. Notice that the latter relationship would not be discovered by using the plain Cosine similarity measure.

For each couple of search results vectors (s_a,s_b) $a,b = 1...m$ we compute the corresponding element of the RSSF equivalence matrix $E \in \Re^{m \times m}$, whose elements e_{ab} are defined as follows:

$$e_{ab} = \begin{cases} 1 & \forall a = b \\ S(\pi_a,\pi_b) & \forall a \neq b : S(\pi_a,\pi_b) \geq \alpha \\ 0 & otherwise \end{cases} \quad (14)$$

Once E is calculated, the primary connectivity graph $G=(V,E)$ is built (step 3). Each vertex v_i of G corresponds to an RSS item. Two vertices v_a and v_b are connected from v_a to v_b if the corresponding element $e_{ab} \in E$ is greater than α. The α-cut value guarantees that every pair of connected RSS items (π_a,π_a) has a semantic relevance of at least α. Figure 4 shows an example of the construction of the graph G for various values of the threshold α.

Starting from G, we build the set of all disconnected sub graphs $d = \{d_1,...d_M\} \subseteq G$(step 4). For example, from Figure 4(a), it would be $d = \{(\pi_1,\pi_2,\pi_3,\pi_4), (\pi_5,\pi_6)\}$ for $\alpha > 0.2$. Each sub graph d_i corresponds to an aggregation of semantically related RSS items linked according to their relationships. Each vertex of d_i is labelled according to the title of the corresponding newspaper article, and the whole sub graph is represented by the title of the most representative newspaper article (i.e., the item's title whose sum of the row elements of E is maximised). As an example, Figure 5 illustrates an aggregation, taken from the ones actually detected by our prototype. As shown in the figure, arrows between two boxes represent the discovered vector similarity condition between the source box and the target box, as per the intrinsic asymmetric nature of the similarity measurement of Equation (11). The representative element is the green one.

[1]The Lucene score of query q for document d is defined in http://hudson.zones.apache.org/hudson/job/Lucene-trunk/javadoc//org/apache/lucene/search/Similarity.html

	π_1	π_2	π_3	π_4	π_5	π_6
π_1	1	0.82	0.53	0	0	0
π_2	0.85	1	0	0	0	0
π_3	0	0	1	0.6	0	0
π_4	0	0	0.32	1	0	0
π_5	0	0	0	0	1	0.9
π_6	0	0	0	0	0.28	1

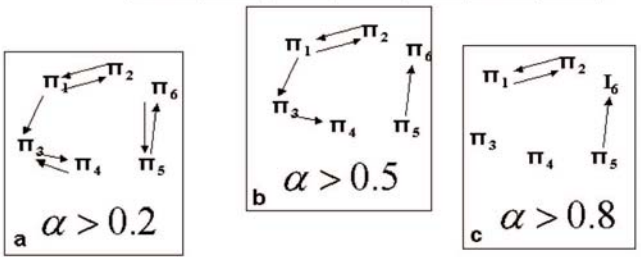

Figure 4. Example of the equivalence matrix between the set of RSS items $\{\pi_i\}_{i=1}^{6}$ and the corresponding connectivity graph for three values of α.

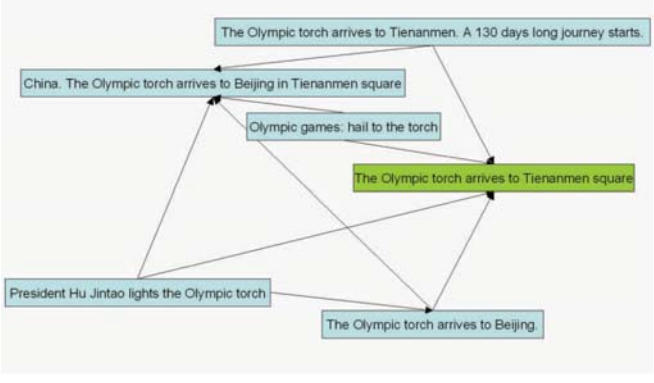

Figure 5. Example of aggregation of semantically related RSS items linked according to their relationships.

The multimedia dossier is finally constructed by taking, for each item $\pi_j \in d_i$, $j = 1...|d_i|$, the set of all the news stories N_k that comply with the following condition:

$$s_{jk} \geq \eta \qquad (15)$$

where s_{jk} is the affinity of the news story N_k to the RSS item π_j (as computed in the first step of the clustering procedure), and η is a fixed threshold. For each multimedia dossier, a text document, which includes both the RSS items' titles and description phrases and all the news stories transcriptions constituting the dossier, is generated. Finally, all such documents are indexed by Lucene and made accessible through the MAi repository (see Figure 1).

2.3 Multimodal Navigation Service

The system supports both simple queries (e.g., one or more search keywords) as well as more advanced queries (e.g., weighted queries, Boolean operators) for searching and retrieving the aggregations. As a simple example, Figure 6 shows an example for the query *"garbage AND Naples"*. To facilitate the results visualisation, the system provides a browsable Web page showing the ranked results. For each retrieved dossier, the system not only lists the basic information, i.e. the dossier's title, score and last update timestamp, and it provides the links for the included news stories and newspaper

Figure 6. Example of the search results Web page for the query *"garbage AND Naples"*

articles. Additionally, as the search results page is provided as an RSS feed, users can subscribe to the submitted query, and automatically receive a notification when the results page is modified, i.e. when either an already included dossier is updated or a new one is discovered.

3. Experimental Evaluations

This section presents the experimental evaluation of all the parts constituting our prototype system. The system was run for seven months, ranging from the end of November 2007 to the beginning of June 2008. In this period of time, we collected about 88,280 online articles and 23,940 news stories, resulting from the segmentation of 3,670 newscast programmes. The online articles were downloaded from 95 RSS feeds supplied by 16 online newspapers and press agencies Web sites. The newscasts were acquired from the daily programming of seven national TV channels. We set $\alpha = 0.8$ in Equation (14) and $\eta = 0.5$ in Equation (15), resulting in a total of 4,187 multimedia dossiers.

3.1 DTV Stream Processing Performance

The TV programme detection and segmentation task is performed acquiring live streams from seven major national channels 24 hours/day and 365 days/year. The system elaborates 16 programmes/day concentrated around the main editions of the newscasts (around 2 p.m. and around 8 p.m.), resulting in about 10 hours per day of elaborated audiovisual material. The total elaboration time on average is about 3.74 times the programme duration, that is normally a newscast of half an hour duration is searchable and retrievable with all its components after less than 2 hours after its end.

We tested the programme detection technique described in Section 2.1.1 in two distinct experiments. In the first experiment, 11 different reference clips had to be identified in a data set constituted by 782 clips taken randomly from daily television schedules. The second experiment consisted in detecting the starting and ending jingles of 7 distinct newscast programmes (total 14 clips) in a continuous flow of acquired audiovisual material. In the first experiment, the measured average detection accuracy of the process was 0.80, while recall was 0.87. In the second experiment reached precision was 1.00, while recall was 0.90, considering as detected items only programmes for which both starting and ending jingles had been identified.

We tested the news story segmentation algorithm described in Section 2.1.2 on about 40 hours of material, for which all true story boundaries were manually identified. To assess the system performance we used an alignment measurement taking into account starting boundaries and ending boundaries with different weights, as well as considering missing material as having more impact than extra material on the measurement. The obtained average precision and recall were 0.76 and 0.73, respectively.

3.2 RSSF Stream Processing Performance

To test the effectiveness of the RSSF stream processing chain, we set up a pool of 25 expert users taken from the employees of our organisation. Each user was asked to evaluate the *consistency* of the generated multimedia dossiers. To this end, we designed and implemented an HTML Web interface, showing a list of dossiers, randomly selected among the available ones. For each dossier, the following three markers were asked to be evaluated, using a judgement scale from 1 (poor) to 5 (excellent):

1. For the whole dossier, assign a semantic cohesion index (I_1) that reflects the overall strength of the RSS and broadcast news streams aggregation;
2. For each included RSS item, assign a consistency index (I_2) to the concept expressed by the dossier;
3. For each included news story, assign a consistency index (I_3) to the concept expressed by the dossier.

In addition, we asked to choose a title T_i *from* among the ones belonging to the aggregated online articles, and to evaluate the representativeness of the chosen title (I_4) on the same judgement scale, w.r.t. the concept expressed by the whole dossier. The effectiveness of the title selection strategy was then evaluated by means of the following indices:

1. The ratio between the number of correctly selected titles (i.e., the number of dossiers for which the system-chosen title agrees with the user-chosen title) and the total number of assessed dossiers (I_5);
2. The average representativeness of the correctly selected titles (I_6);
3. The average representativeness of the wrongly selected titles (I_7);
4. The average representativeness of the user-selected titles (I_8).

The evaluations were collected from the end of March 2008 to the beginning of June 2008, obtaining a total number of 651 evaluated dossiers. Table 1 reports the values obtained for the indices I_1 to I_8 in terms of average, standard deviation and confidence interval. Overall, the system shows an outstanding performance, getting in most cases a score of either 4 or 5 in most of the indicators that was chosen for assessment. Poor performances for I_5 seem to indicate that the algorithm used by the system to choose the dossier's title needs to be further improved. However, as $I_6 > I_8$ and $\sigma_6 < \sigma_8$, we can state that the titles chosen by the system are more representative than the others. Furthermore, $I_8 \cong I_7$, indicating that when the title automatically selected are wrong, they are still significantly relevant to the concept expressed by the dossier.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
Value	4.23	4.65	4.24	4.66	0.31	4.85	4.63	4.66
Std. Dev.	0.85	0.89	1.17	0.70	n.a.	0.62	0.77	0.70
C.i. start	4.19	4.62	4.20	4.59	n.a.	4.84	4.47	4.54
C.i. end	4.35	4.68	4.28	4.70	n.a.	4.89	4.69	4.70

Table 1. Accuracy indices of the multimodal aggregation service

4. Conclusions

Efficient large-scale implementation of multimedia indexing systems is an hard task, since operating conditions of real systems may put very stringent requirements on the overall performances, and real-life environments are often quite diversified w.r.t. prototype ones. As users may want to browse through large amount of news contents without spending much time in searching for the desired information, multimedia news management applications must provide search and retrieval services in a timely and efficient manner.

To this end, this article presented and discussed the implementation of a large-scale automatic acquisition, segmentation and indexing of broadcast news content. The overall system performances are encouraging, and the accuracy of the programme boundary detection and segmentation are at a good level of maturity.

In addition, this article introduced a novel method for aggregation of multimodal sources of information, based on a semantic relevance function acting as a kernel to discover the semantic affinities of heterogeneous information items, and on an vector projection similarity principle on which semantic dependency graphs among information items can be constructed. The generality of the proposed method allows its application to a wide collection of real cases, and in this article we presented an application of it in the area of multimodal news aggregation. We developed a prototypal system, which has been evaluated by collecting daily television newscasts from 7 major Italian broadcasters and 95 RSS news feeds from online websites for a period of about 8 months. Obtained results are very encouraging and demonstrate the robustness and effectiveness of our method.

Future developments will regard both the theoretical foundations and the practical outcomes of our technology. From the theoretical perspective, models of integration of more than two information items sets will be investigated. On the implementation side, to further ensure a full-potential exploitation of our system in industrial contexts, future work will also regard both the improvement of the general performances in terms of elaboration latency and in terms of the accuracy of most critical processing blocks of the system, based on the performances presented in this work.

References

[1] RSS Specifications, www.rss-specifications.com/

[2] OpenLogos Machine Translation, http://logos-os.dfki.de/

[3] TreeTagger – a language independent part-of-speech tagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4] Brugnara, F., Cettolo, M., Federico, M., Giuliani, D. (2000). A system for the segmentation and transcription of Italian radio news. In *Proc. of RIAO, Content-Based Multimedia Information Access,* Sept. 2000.

[5] Basili, R., Cammisa, M., Donati, E. (2005). Ritroverai: A web application for semantic indexing and hyperlinking of multimedia news. In *Proc. of International Semantic Web Conference*, pages 97-111, June 2005.

[6] Banerjee, S., Ramanathan, K., Gupta, A. (2007). Clustering short texts using Wikipedia. In *ACM SIGIR Conf. on Research and Development in Information Retrieval*, July 2007.

[7] Chua, T., Chang, S., Chaisorn, L., Hsu, W. (2004). Story boundary detection in large broadcast news video archives: Techniques, experience and trends. In *ACM Multimedia 2004*, Oct. 2004.

[8] De Santo, M., Percannella, G., Sansone, C., Vento, M. (2006). Unsupervised news video segmentation by combined audio-video analysis. *In: Proc. of the Intl. Work. on Multimedia Content Representation, Classification and Security (MRCS)*, pages 273-281, Sept. 2006.

[9] Hoashi., K. (2004). Shot boundary determination on MPEG compressed domain and story segmentation experiments for TRECVID 2004. In *Proc. Of TRECVID Workshop 2004*, Feb. 2004.

[10] Hsu, W., Changy, S.F., Huangy, C.W., Kennedyy, L., Linz, C.Y., Iyengar, G. (2004) Discovery and fusion of salient multi-modal features towards news story segmentation. *In Proc. of Storage and Retrieval Methods and Applications for Multimedia*, pages 244-258, Jan. 2004.

[11] Huang, W., Webster, D. (2004). Intelligent RSS news aggregation based on semantic context. In *ACM SIGIR Workshop on Information Retrieval in Context*, pages 40-42, July 2004.

[12] Kraaj, W., Smeaton, A., Over, P. (2004) TRECVID 2004: An overview. In *Proc. of TRECVID Workshop 2004*, Feb. 2004.

[13] Li, X., Yan, J., Deng, Z., Ji, L., Fan, W., Zhang, B., Chen. Z. (2007). A novel clustering-based RSS aggregator. In *16th Int. Conf. on World Wide Web*, pages 1309-1310, May 2007.

[14] Delèglise, P., Estève, Y., Meignier, S., Merlin, T. (2005). The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news. In *In Proc. of Interspeech'05*, Sept. 2005.

[15] Pickering, M.J., Wong, L., Rueger, S.M. (2003). Anses: Summarisation of news video. In *Proc. of International Conference on Image and Video Retrieval (CIVR)*, pages 481-486, Jan. 2003.

[16] Quènot, G. M., Mararu, D., Ayache, S., Charhad, M., Besacier, L. (2004). CLIPS-LIS-LSR-LABRI experiments at TRECVID 2004. In *Proc. of TRECVID Workshop 2004*, Feb. 2004.

[17] Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. of the Int. Conf. on New Methods in Language Processing*, Sept. 1994.

[18] Tamura, H., Mori, S., Yamawaki, T. (1978). Texture features corresponding to visual perception. *IEEE Trans. on Systems, Man and Cybernetics*, 8(6):460-473, 1978.

[19] Volkmer, T., Tahahoghi, S.M.M., Williams, H.E. (2004). RMIT university at TRECVID 2004. In *Proc. of TRECVID Workshop 2004*, Feb. 2004.

[20] Xu, C., Wang, J., Lu, H., Zhang, Y. (2008). A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video. *IEEE Transactions on Multimedia*, 10(3):421-436, April 2008.

[21] Xu, H., Chua, T.S. (2006). Fusion of AV features and external information sources for event detection in team sports video. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):44-67, Feb. 2006.

[22] Zhai, Y., Chao, X., Zhang, Y., Javed, O., Yilmaz, A., Rafi, F. (2004). University of Central Florida at trecvid 2004. In *Proc. of TRECVID Workshop 2004*, Feb. 2004.

[23] Zhang, Y., Xu, C., Rui, Y., Wang, J., Lu, H. (2007). Semantic event extraction from basketball games using multi-modal analysis. In *IEEE Int. Conf. On Multimedia and Expo*, pages 2190-2193, July 2007.

[24] Montagnuolo, M., Messina, A. (2009). Parallel Neural Networks for Multimodal Video Genre Classification. In *Multimedia Tools and Applications*, 41(1):125-159, Jan. 2009.

[25] Apache Lucene project, http://lucene.apache.org/

[26] Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* 35: 99–109.

Authors Biographies

Dr. Maurizio Montagnuolo received his Laurea degree in Telecommunications Engineering from the Polytechnic of Turin in 2004, after developing his thesis at the RAI Research Centre. Currently. In 2008 he received his Ph.D. in "Business and Management" at the University of Turin, in collaboration with RAI, and supported by EuriX S.r.l., Turin. His initial contributions to computer science were in the area of artificial intelligence, specifically in the semantic classification of audiovisual content. In particular, he was involved in research projects on automatic classification and characterisation of television genre. His current research interests are mostly addressed in the context of Web data mining and multimedia data mining. He is also co-*author of several scientific* works published on Journals or in International Conferences in this subject area.

Alberto Messina began his collaboration as a research engineer with RAI in 1996, when he completed his MS Thesis in Electronic Engineering (at Politecnico di Torino) about objective quality evaluation of MPEG2 video coding. After starting his career as a designer of RAI's Multimedia Catalogue, he has been involved in several internal and international research projects in the field of digital archiving, with particular emphasis on automated documentation, and automated production. His current interests are ranging from file formats and metadata standards to the domain of content analysis and information extraction algorithms, where he now concentrates his main focus. Recently, he has started promising research activities concerning semantic information extraction from the numerical analysis of audiovisual material, particularly in the field of conceptual characterisation of multimedia objects, genre classification of multimedia items, automatic editorial segmentation of TV programmes. He is also author of technical and scientific publications in this subject area. He has extensive collaborations with the local University of Torino – Computer Science Department, which include common research projects and students' tutorship. To complete his scientific formation, he has recently decided to take a PhD in the area of Computer Science. He is active member of several EBU projects including P/TVFILE, P/MAG and P/CP, chairman of the P/SCAIE project dealing with automatic metadata extraction techniques. He is currently working in the EU PrestoSpace project in the Metadata Access and Delivery area. He has served as Programme Committee Member in a Special Track of the 10th Conference of Italian Association of Artificial Intelligence, and in the Workshop on Ambient media Delivery and Interactive Television.

Roberto Borgotallo graduated in Telecommunication Engineering at *Politecnico di Torino* in 1999. Since 2001, he has been working for RAI - Radiotelevisione Italiana at the R&D department (*Centro Ricerche e Innovazione Tecnologica*) in Turin. Initially, he was involved in several projects gravitating around the RAI multimedia catalogue, called CMM, regarding metadata ingestion and transformation. More recently, Mr Borgotallo has been working in a team that is developing an automatic metadata extraction platform which is actually used extensively in RAI for experimental purposes, and even for real in the production evironment. His major professional interests are metadata and essence transformation, system integration and workflow management.