ViRMA: Virtual Reality Multimedia Analytics at Video Browser Showdown 2022

Aaron Duane and Björn Þór Jónsson

IT University of Copenhagen
aadu@itu.dk, bjth@itu.dk

Abstract. In this paper we describe the first iteration of the ViRMA prototype system, a novel approach to multimedia analysis in virtual reality, that is inspired by the M^3 data model. In this model, media is mapped into a multidimensional space, based on its metadata. ViRMA users can then interact with the media collection by dynamically projecting the metadata space to the 3D virtual space, through a variety of interactions, and exploring the resulting visualisations.

Keywords: virtual reality · human-computer interaction · multimedia analytics

1 Introduction

In recent years, increasing emphasis has been placed on interactive analysis by users in various multimedia application domains, as it has become clear that to satisfy diverse and dynamic information needs, effective collaboration between human and machine is necessary [4]. This calls for combining sophisticated multimedia analysis, scalable data management, and interactive visualisation into a single system that supports the user's interactive analysis of a media collection [7].

At the same time, hardware for interacting with data in virtual reality has been improving rapidly, and has reached the point where quality interactions are possible with affordable hardware. Past research has suggested that VR is highly valuable due to its immersive quality, the degree to which it projects stimuli onto the sensory receptors of users, and that it will lead to more natural and effective human-computer interfaces [5]. In this paper we present the first iteration of the ViRMA (Virtual Reality Multimedia Analysis) prototype.

The foundation for the ViRMA prototype is the M^3 model [1] which is in turn based on the merging of concepts from business intelligence, such as analytical processing (OLAP), multidimensional analysis (MDA), and faceted browsing. ViRMA intends to utilise a visualisation paradigm which relies explicitly on the effective representation of multimedia data in 3D virtual space. With the ViRMA prototype, we aim to consider the impact of collection scale on VR interfaces to multimedia analytics, and use interactive competitions such as VBS and LSC [2] as a platform to evaluate our approach.

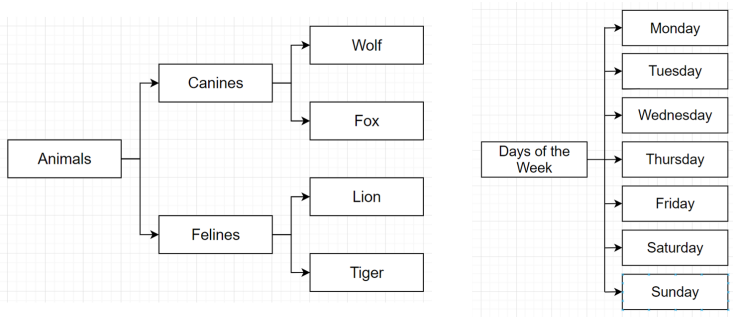

Fig. 1. Example of a hierarchy (left) and a tagset (right)

2 The VBS Dataset and M^3

The M^3 data model (pronouced "emm-cube") considers multimedia objects to reside in a multi-dimensional metadata space, and then provides support for exploring that space via operations to project it down to 3D space. In the context of the VBS collection, the media objects are extracted keyframes, taken from the collection's video sequences, which results in a dataset of several million images. In addition to the metadata already provided in the video collection, we have extracted semantic features from these keyframes using the ImageNet Shuffle [3], a deep neural network using the ResNeXt-101 architecture [6]. For each of the 1 million images, the 5 highest-scoring concepts are retained as tags, resulting in several thousand unique tags. Since the tags extracted by ImageNet Shuffle directly correspond to a subset of the WordNet database, we created a large hierarchy containing every distinct tag using the WordNet Python API.

3 System Description

The core components of ViRMA system can be described as falling under two main categories; components which support the generation of user queries, and components which support the visualisation and exploration of the data produced by these queries. We refer to these categories respectively as *query generation* and *data projection* within the context of the ViRMA system. In this section we will describe each of the components in these categories and how the user might interact with them in order to complete a typical VBS task.

3.1 Query Generation

To begin, the metadata extracted from the VBS collection is organised semantically into tags, tagsets, and hierarchies (see Figure 1). These serve the underlying M^3 model but, from the perspective of the user, these serve as potential filters which can be applied to their current query or browsing state.

It is important to note that within ViRMA, these filters can be applied to the data in two fundamental ways. The most obvious approach is to locate a tag, tagset, or hierarchy of interest, and simply apply it as a direct filter on the dataset. This will reduce the entire set of potential results to those which are tagged by that filter. In the case of a tagset or hierarchy, this will include all of their children. The other fundamental way of applying a filter in ViRMA, is to visualise it on a spatial axis in the virtual environment. We refer to this as projecting the filter as a dimension and will explore it in more detail in the *data projection* section. With so many potential filters available in ViRMA, including the different methods by which they can be applied, it is imperative that a user can effectively navigate and browse this metadata before deciding which are most appropriate to use and how they should be applied.

Browsing or searching for a tagset is comparatively straightforward as each tagset only contains a single group of semantically associated tags. Hierarchies, however, can contain any number of tags at varying depths of association and can quickly become cumbersome to navigate when searching for a specific tag or tagset in the hierarchy. One option is to restrict the depth and complexity of hierarchies so they are easier to navigate, but this reduces the utility of the hierarchy once it is projected as a dimension in the virtual space. To maintain the benefits of more semantically detailed hierarchies and yet mitigate their associated detriments, we introduced the concept of a *dimension explorer* to the ViRMA system which we will now describe.

Dimension Explorer The dimension explorer is a dedicated user interface element which can be opened at any time within the virtual space. Once loaded, it is populated with a list of all tagsets and all hierarchies (at their topmost level) which are currently available for the dataset. From this list, the user can drill down into any tagset or hierarchy and their children before deciding to apply them as either a direct filter or project them as a dimension in the virtual space. When viewing the contents of a hierarchy in the dimension explorer, the user's depth in the hierarchy is contextualised by also displaying it's parent and children, if any exist.

As has been noted, in situations with deep or complex hierarchies, this approach can become tedious when the user cannot locate a specific tag or tagset they are searching for. To address this, the dimension explorer contains a search input which enables the user to search for specific tags or tagsets within their respective hierarchies. Selecting this search input will open a virtual keyboard which the user can interact with inside the virtual space using their wireless controllers. By submitting whole or partial tag names, the dimension explorer is loaded with any matching or related tags and tagsets whilst continuing to visually preserve their context within the hierarchy. This empowers the user to select whatever depth of a hierarchy that is appropriate for their current query or browsing state to be applied as a direct filter or to be projected into the virtual space.

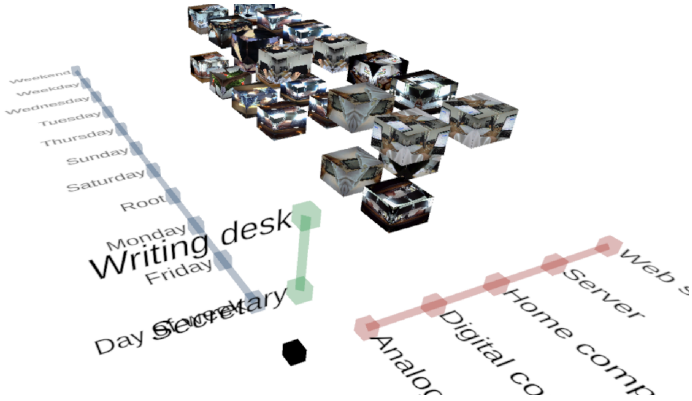

Fig. 2. A browsing state with three projected dimensions

3.2 Data Projection

Now that we have established the primary interaction mechanism by which a user can apply filters to the target dataset, we can begin to explore how this data is visualised within ViRMA's virtual environment. As previously stated, we refer to the concept of applying a filter as a visible dimension in the virtual space as *data projection*. This is to draw a distinction between data or metadata which can be visualised as a dimension on a spatial axis in the environment and data or metadata which can be visualised in other contexts, such as within the dimension explorer. If we consider all the data and metadata organised by the M^3 model as existing in a multi-dimensional space, projecting dimensions in the ViRMA system is the equivalent of taking a specific slice of that multi-dimensional space and mapping it to the three spatial dimensions available in our virtual environment.

Slicing and Dicing When the user first loads into the ViRMA system, the virtual environment is empty and the only interactive elements present are those necessary to support initial query generation, such as the dimension explorer. We refer to the virtual environment within ViRMA as the projection space. Upon selecting a filter the user wishes to project as a dimension, and choosing which axis the user wants to map the dimension to (i.e. X: left/right, Y: up/down, or Z: in/out), the projection space is populated with the relevant axis and a representation of the data along that axis (see Figure 2).

In the context of the M^3 model, projecting more than one dimension is referred to as *dicing* the multi-dimensional space and, from the perspective of the user, provides a data representation that conveys groups of images containing the various alcoholic drinks on each day of the week. The user can continue slicing and dicing the multi-dimensional space by projecting to the remaining third spatial dimension with any appropriate filter that suits their query.

Drilling Down and Rolling Up When a tagset that exists as part of a hierarchy is projected to an axis, the user has the additional option to *drill down* or *roll up*. In the context of drilling down, this would involve drilling into a child of the current tagset and re-populating the projection space with that child's children on the axis instead. For example, with the "alcohol" tagset, a user might want to drill into "spirits" on the axis. The "spirits" tagset might contain children such as "whiskey", "vodka" or "rum" and upon drilling down, they would replace the parent tagset which was previously applied to the axis.

For rolling up, this naturally produces the opposing effect and involves re-populating the axis with the parent of the current tagset, and subsequently any siblings of that parent on that level. For example, if we rolled up from the children of the "alcohol" tagset, we would populate the axis with the "beverage" tagset, and will see "alcohol" and any of its siblings, such as "soda", "tea" or "coffee" now on the axis. Drilling down and rolling up can be accomplished easily in the ViRMA system by targeting a specific axis with one of the wireless controllers and selecting the appropriate contextual action which is displayed to the user.

Pivoting Once a dimension has been applied to an axis in the projection space, it is important to understand that this dimension can be removed from the axis in two fundamentally different ways. The first method is the user can simply clear the dimension entirely, and all filters that were associated with the dimension are removed from the browsing state, increasing the amount of potential results in the current query. The second method is to replace the dimension on an axis whilst maintaining that dimension's filters on the browsing state. This is referred to as *pivoting* in the context of the M^3 model and is the equivalent of using the dimension explorer to apply a number of tags or tagsets as direct filters before projecting a different tagset as a dimension to one of the axes in the projection space. It is imperative that this concept is effectively conveyed to the user in the ViRMA system as it is fundamental in the user's understanding of the current browsing state within the projection space.

Cell and Timeline Exploration Once the user has sufficiently refined their query using the aforementioned techniques, it is likely that they will want to explore the image contents of an individual cell in the data projection space. This can be accomplished at any time by pointing one of the wireless controllers at the relevant cell and selecting the contextual option that appears. This will temporarily reload the projection space with all of the images contained in that cell. Furthermore, while browsing a cell's contents, if the user wishes to view any individual result in the context of the wider lifelog, they can select the image via a contextual interaction and it will load a timeline above the cell's contents displaying the image as it appeared in the lifelog. The user can then scroll left in this list to move backwards in time and right to move forwards in time. Finally, the user may return to the original projection space by selecting the appropriate button on either controller.

4 Conclusion

The ViRMA system prototype is the first iteration of a novel virtual reality multimedia analysis platform based on the M^3 model. In this paper we have attempted to describe the system with respect to its underlying data model. It is our hope that we can evaluate our approach via VBS 2022 to serve as a benchmark against other multimedia analytics systems.

Acknowledgement: *This work was supported by MCSA-IF grant 893914.*

References

1. Gíslason, S., Jónsson, B.Þ., Amsaleg, L.: Integration of Exploration and Search: A Case Study of the M^3 Model. In: Proc. MMM. pp. 156–168 (2019)
2. Gurrin, C., Jónsson, B.Þ., Schöffmann, K., Dang-Nguyen, D.T., Lokoč, J., Tran, M.T., Hürst, W., Rossetto, L., Healy, G.: Introduction to the fourth annual lifelog search challenge, lsc'21. In: Proc. ACM ICMR. Taipei, Taiwan (2021)
3. Mettes, P., Koelma, D.C., Snoek, C.G.: The ImageNet shuffle: Reorganized pre-training for video event detection. In: Proc. ACM ICMR. pp. 175–182 (2016)
4. Seebacher, D., Häußler, J., Stein, M., Janetzko, H., Schreck, T., Keim, D.A.: Visual analytics and similarity search: concepts and challenges for effective retrieval considering users, tasks, and data. In: Proc. SISAP. pp. 324–332 (2017)
5. Slater, M., Wilbur, S.: A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. Presence: Teleoperators and Virtual Environments **6**(6), 603–616 (12 1997)
6. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proc. CVPR (2017)
7. Zahálka, J., Worring, M.: Towards interactive, intelligent, and integrated multimedia analytics. In: Proc. VAST. pp. 3–12 (2014)