IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Utility Patent Application (Provisional)

TITLE: NOVEL METHOD FOR THE INTEGRATION OF ARTIFICIAL INTELLIGENCE AND COMPUTER VISION BASED DEEPFAKE ALGORITHMS FOR THE MANIPULATION OF MOUTH AND TEETH IN VISUAL MEDIA FOR THE ENHANCEMENT OF DENTAL CARE, DENTAL HEALTH, DENTAL EDUCATION, COSMETIC APPEARANCE, DENTAL COSMETICS, LIP COSMETICS, MOUTH IMAGERY ALTERATION, MOUTH VIDEO ALTERATION AND THE LIKE

INVENTOR: George Alexander Davila

FIELD OF THE INVENTION

[0001] The present innovation concerns novel algorithms and collections of interrelated sub-processes of such algorithms which utilize modern methods in artificial intelligence (AI), deep learning, machine learning, and computer vision for the manipulation of images, videos, and other visual media of

human teeth and human mouths. More particularly, the present invention relates to artificial intelligence algorithms designed to allow for the easy and rapid modification and/or enhancement of dental imagery such that a dentist, dental patient, and/or layperson may edit dental, mouth, or tooth/teeth imagery (or alternative visual representations such as, but not limited to, video), allowing for the enhancement of patient-dentist coordination in dentistry, the enhancement of dental hygiene education and outreach, general cosmetic enhancement, as well as number of other applications. One such manifestation and exemplification of such algorithms may be the use of artificial intelligence(s) to autonomously replace one individual's mouth with the mouth of another individual in order to show the impact of a particular dental procedure.

[0002] The present disclosure and discussion is to be considered as an exemplification of the invention and a use case of note, and is not intended to limit the invention to the specific embodiments or applications illustrated by the figures, descriptions, or discussions above or below.

BACKGROUND

[0003] Dental care, like many other industries, has seen a rise in digital and e-commerce centered interactions. Consumers today have the option of ordering dental care products – from toothpastes to braces, amongst many other products – almost entirely online. Many other interactions have shifted to low-touch interactions, that is to say the contemporary consumer might be able to interact with their dental care providers mostly online with the occasional in-person consultation when health and/or standards of care and/or regulations require such interaction.

[0004] Patients often fear dental procedures and/or other mouth based procedures for their potential impact on ones appearance, as negatively impactful cosmetic procedures may negatively affect one socially, mentally or otherwise. Such concerns raise the need for methods by which individuals can be shown potential outcomes in a rapid, and ideally realistic, manner.

[0005] Patients often lack proper education and insight about dental care. While dentists may be aware of the negative impacts of some dental procedure, communicating such impacts to each and every patient is in itself a challenge.

[0006] To further elucidate and expand upon the concern(s) raised in parts [0004] & [0005] we may consider the case of children. It is quite common for children to seek to forgo proper dental care, from not brushing their teeth to eating too much sugar to attempting to avoid dentist visits all together. Children are particularly susceptible to misunderstanding or ignoring relatively abstract and often verbose dental recommendations. Visual aids, such as those showing images of positive and negative dental outcomes, are often useful in educating children in such matters. Photorealistic overlays of potential dental outcomes over a child's own face in real time may help more greatly elucidate such matters.

[0007] Considering the case of dental veneers we may observe a set of tasks which are time consuming, costly, and potentially injurious. Common practice, even in the modern era, is to shave and chisel ones natural teeth down simply to place on a set of trial veneers, often before the proper set is ever even crafted. This trial set will likely be the first time a patient may explore and examine the full cosmetic impact of such veneers. It is not uncommon for people to experience malalignment between an imagined cosmetic outcome described by a professional cosmetic service provider and the real outcome as perceived by the

person themself. The layperson might experience such discrepancies in cosmetically focused procedures & procurements such as haircuts, buying clothes online, amongst many other such instances. In the instance of dental veneers, however, one's natural teeth are most often permanently altered before the patient can ever truly see the results. With hundreds of thousands to millions of Americans getting veneers at a cost ranging from hundreds to thousands of U.S. dollars per tooth, such discrepancy issues can have massive impacts on society. Such less-than-desirable cosmetic work may have impacts on patient finances and patient mental health, as they may need to spend more to repair the work and may suffer unnecessary hardship due to the financial burden as well as the sub-par cosmetic work.

[0008] Dentists may have difficulties in conveying the impact of some dental procedure or lack thereof through no fault of their own. The layperson might not be expected to fully comprehend the impact of any number of potential dental outcomes. Nor can every dentist be reasonably expected to properly be able to convey every possible impact, as it is in the interest of society that individuals be made dentists because they can produce optimal health outcomes rather than for their skill in patient communication. It might also be the case that a dentist might have to negatively impact the cosmetic appearance of a patient's teeth in order to

achieve an optimal health outcome. In such cases, a system whereby a dentist might do so in such a way as to minimize patient mental distress by coordinating and communicating the cosmetic alterations, and the nature of such alterations, might be considered desirable.

[0009] Similar difficulties may arise in other forms of mouth alteration, such as—but not limited to—in plastic surgery, piercing, or lip injections. The permanent and/or potentially injurious nature of these applications renders them of particular note, and heightens the need for the lessening of these difficulties.

Of relation to plastic surgery applications are instances of mouth/lip deformation or destruction. Cleft palate, for instance, is expected to occur in approximately every 1 of 800 people born in the United States and approximately every 1 of 600 people born worldwide according to estimates. In many such instances the individual to be operated upon or their caretaker may lack knowledge of the process or even of modern medicine itself in the case of treatment in particular global regions. A method to visually display expected outcomes on and/or to the individual in question may therefore be highly desirable and highly effective in conveying the reasons for treatment. Partial or total destruction of the lips may occur in ailments

or injuries such as those suffered by extreme burn victims. In such a scenario the lips may have to be reconstructed nearly entirely from nothing, or in the extreme case of a face transplant may be transferred from an extant individual. In such scenarios it may be highly desirable to preview the potential replacement lips so that the patient may be most ably equipped for proper recuperation and reduction of anguish while potentially minimizing the need for future reconstructive surgery.

[0010] Therefore, a need exists in the field for novel technologies which allow a patient to visually and photo realistically preview the potential impacts of dental work—or other forms of mouth alteration—on their face. The present methods allow the consumer to gain greater insight into the impact of potential care via digital means, thus addressing issues, problems, and trends discussed in [0003], [0004], [0005], [0006], [0007], [0008], [0009].

DETAILED DESCRIPTION OF THE INVENTION

[0011] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention. As used herein, the term "and/or" includes any and all combinations of one or more of

the associated listed items. As used herein, the singular forms "a," "an," and "the" are intended to include the plural forms as well as the singular forms, unless the context clearly indicates otherwise.

It will be further understood that the terms "comprises" and/or "comprising," when used in this specification, specify the presence of stated features, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, steps, operations, elements, components, and/or groups thereof.

[0012] Unless otherwise defined, all terms (including technical and scientific terms) used herein have the same meaning as commonly understood by one having ordinary skill in the art to which this invention belongs. It will be further understood that terms, such as those defined in commonly used dictionaries, should be interpreted as having a meaning that is consistent with their meaning in the context of the relevant art and the present disclosure and will not be interpreted in an idealized or overly formal sense unless expressly so defined herein.

[0013] In describing the invention, it will be understood that a number of techniques and steps are disclosed. Each of these has individual benefit and each can also be used in conjunction with one or more, or in some cases all, of the other disclosed techniques. Accordingly, for the sake of clarity, this description will refrain from repeating every possible combination of the individual steps in an unnecessary fashion. Nevertheless, the specification and claims should be read with the understanding that such combinations are entirely within the scope of the invention and the claims.

[0014] The term deepfake is herein intended to refer to the autonomous manipulation of human depiction(s) in photos, images, and/or videos via the utilization of neural network based algorithms, facial recognition algorithms, or other artificial intelligence algorithms, or some combination(s) of such algorithms as well as the results of such manipulation(s). The term deepfaking is herein intended to refer to the process(es) of, and involved in, creating deepfake media.

[0015] New autonomous mouth deepfake placement and photo editing devices, apparatuses, and methods for deepfaking the human mouth are discussed herein. In the following description, for purposes of explanation, numerous

specific details are set forth in order to provide a thorough understanding of the present invention. It will be evident, however, to one skilled in the art that the present invention may be practiced without these specific details.

[0016] The present disclosure is to be considered as an exemplification of the invention, and is not intended to limit the invention to the specific embodiments illustrated by the figures or description below.

[0017] A major innovation present is the utilization of artificial intelligence and related technologies for the tasks at hand. This allows for the instant and autonomous replacement of the human mouth. An equivalent manual photo editing process would incur time and/or labor costs while still not allowing users to view these results in real time.

[0018] Various segmentation methods are depicted and/or discussed in FIG. 2, FIG. 3, FIG. 4, and FIG. 5. Segmentation, while an important tool, is only a part of our algorithm. The present algorithm is versatile enough to use a variety of segmentation methods interchangeably. Three specific segmentation methods are herein referred to as polygonal segmentation, landmark segmentation, and full-face

segmentation, and are depicted by FIG. 2, FIG. 4, and FIG. 5, respectively. Furthermore, mouth-only segmentation may refer to both polygonal segmentation and landmark segmentation, as both may be used to segment only the mouth, in contrast to full-face segmentation of FIG. 5. Each segmentation method may have specific attributes, advantages, or disadvantages associated with it. It is therefore helpful to preserve optionality in the choice of segmentation method so that we may optimize such choices for specific use cases. When referring to segmentation without any contextual clarification, we may be taken to be referring to the use of any such segmentation method. In particular, segmentation as discussed in the algorithmic process flow diagram of FIG. 3 may be taken to refer to the use of any such segmentation method. Our focus on these segmentation methods is only intended to more clearly elucidate the nature of the algorithm, and is not intended to limit the invention to the specific embodiments or applications of segmentation methods illustrated by the figures, descriptions, or discussions above or below.

[0019] The present invention will now be described by referencing the appended figures representing preferred embodiments. FIG. 1 depicts an example potential process flow for the autonomous replacement of the human mouth. Individual A is the “source” individual from whom a model mouth is taken. Individual B is the target individual, or the subject of the algorithm or an apparatus

on which the algorithm is acting, on whom the model mouth is placed. In following discussions, we will continue the use of this nomenclature convention such that a “source mouth” shall be regarded as an extant mouth. The “source face” shall refer to the face from which the source mouth is extracted. A “user mouth” shall be regarded as the original mouth of the aforementioned user of the algorithm. The “user face” shall refer to the face from which the user mouth is extracted. We also assert the user to be the individual on whom the final deepfake mouth is placed. Any third party individual(s) running the algorithm shall be regarded as “operator(s)”. Of note is that the user, as defined herein, is any subject on whom the algorithm is ran. While such a user, as defined herein, is assumed to most often be interacting with the algorithm, and therefore also an operator, it may be the case that they are not and that an extant operator is the only individual interacting with the algorithm. In the instance that the user is not an operator we may alternatively refer to them as the “subject”, their original mouth may be referred to as the “subject mouth” and the face from which that mouth is extracted as the “subject face”. The user is therefore always a subject and always an operator as defined herein. An operator is only the subject if they are also the user. A user may also be the source. A subject may also be the source in the instance where an operator is not the user. An operator who is not they subject may even select themselves as a source, deepfaking their mouth unto another individual. A real

world use case might entail a dentist showing a patient what that patient's teeth might look like after a whitening. In this use case both the dentist and patient would be operators, the patient would also be the subject and a user as their face is the one acted upon. If the dentist chooses to edit the patient's own teeth to make them whiter then the patient is would also be the source. If the dentist then chose to edit the mouth of decease United States President Abraham Lincoln, for instance, then both the patient and dentist would be non-user operators and Abraham Lincoln the non-user subject. If the dentist opted to place their own mouth on Abraham Lincoln, then the dentist would become the non-user source.

The term "original source mouth" shall be used to refer to the source mouth prior to any alteration and the term "final source mouth" shall be used to refer to the mouth placed unto the user's face. Notable is the case of the equivalence of the user mouth and the source mouth. It may be the case that the user simply wishes to edit – through processes to be described below – their own mouth see what the resultant edit looks like. In such a scenario the user would be both individual A and individual B as depicted in FIG. 1. The terms user mouth and source mouth may be used interchangeably up to the time when the first edit occurs. After such instance, the edited mouth will be referred to as the source mouth and the user's original mouth will continue to be referred to as the user mouth. As defined above, the term original source mouth would be equivalent to the user mouth in this instance.

For instance, it may be the case that the user is missing a tooth and – through processes to be described below – proceeds to use the present algorithm to edit their mouth to restore the missing tooth. Let us suppose this only requires a single edit. In such an instance the user's mouth would be the source mouth, with the equivalence only breaking at the point in time when this edit occurs.

In consideration of the real world use cases expected to be the most prevalent, we will frame following discussions in terms of the case of the user – such that the subject is assumed to be an operator – unless otherwise specified.

Three time intervals are denoted by labels 1A, 1B, 1C, with 1A being associated with the earliest time interval, 1B being the middle time step, and 1C being associated with the final time interval. 1A depicts both individuals unaltered. The time interval of 1B depicts the placement of the mouth of Individual A onto Individual B. During this process, all properties of the model mouth are preserved. In the transition from the time interval of 1B to the time interval of 1C the mouth is fitted to better suit the face of individual B. Such processes may take the form of skin tone blending, size readjustment, generative adversarial network based embedding, or a number of other visual adjustments, enabled by either advanced artificial intelligence algorithms or computer vision based algorithms.

[0020] FIG. 2 depicts a potential representation of segmentation of the human mouth. In the instance depicted in FIG. 2 the human mouth is estimated to be contained within the inner boundary labelled 2A, the whitened area contained within this boundary is herein referred to as the segmentation instance mask. The blackened region labelled 2B is referred to as the segmentation background mask. As discussed in [0018] we will refer to this style of segmentation as polygonal segmentation. Polygonal segmentation such as that in FIG. 2 may be autonomously obtained via the use of artificial intelligence algorithms or via the use of a variety of other computer vision techniques, although artificial intelligence algorithms have begun to exceed other computer vision algorithms in both efficacy and efficiency – as applied to the general case – in recent years. We now refer again to FIG. 1. Such a segmentation as depicted in FIG. 2 allows us to extract the region of the mouth from individual A in FIG. 1. Then, by similarly creating segmentation masks for individual B, we may superimpose the mouth of individual A unto individual B. Before such superimposing, but after the extraction of the mouth from individual A, we may apply a multitude of morphing or other processing techniques. Such intermediate processing allows us to: manipulate the source mouth without manipulating the image of either individual, and lower the computational costs associated with any such manipulation.

[0021] FIG. 3 depicts one likely potential algorithmic process flow. Potential steps in the algorithm are denoted (0) through (8). 3A marks the first potential divergence. At 3A we may elect to either edit the source mouth through steps (3) and (4) or both. Such editing will require additional computational resources. This editing could occur in the instance where the user wants to continue using their own mouth throughout the process and sufficient computational resources can be allocated to the task. Step (0) entails the input of user data, which may entail details such as which media source should be edited, the details of the desired result, amount of computational resources to dedicate to the task, amongst any other details which may need to be added. Many of these details may be assumed, set to a predefined default, or determined by the computational resources and hardware which the algorithm can access. To avoid such editing, we may use the mouth of a preselected human model, extracting the mouth from them in step (1). Such models may be selected for features considered visually desirable such as white teeth, and for their similarity to the user, such that a female mouth is placed on a female user and so on. The use of such models does not necessarily forbid the editing of steps (3) and/or (4), such as if a user desired to see most of the model's mouth with slight alterations. We may also opt to cycle between steps (3) and (4), returning to step (3) after a run of step (4). Such an optimization loop may be used to better fit the GAN produced image to the shape

of the deepfaked mouth. Step (4) independently may entail, but is not limited to, tasks such as simplistic tooth whitening, vertically flipping a tooth on either the left or right side to overlay onto its opposite side counterpart, or reshaping the mouth to perform a rough straightening. While not as visually appealing as the more computationally burdensome GANs-produced results, a user may opt to use only step (4) in order to preserve computational resources. Notably step (3) is where our additionally novel mouth-focused GANs may be applied. The later GANs of step (7) makes use of standard face-focused GANs. 3B denotes the second potential divergence. Once the mouth is placed on the user we may either display it to them immediately, perform a simple embedding and lending via lightweight procedures such as color or size adjustment; or we may perform a GANs based embedding and blending by placing the user's face image of step (5) into the latent space of a face-focused GAN and manipulating it within the scope that such face-focused GANs would allow. Skipping any embedding or blending (null case) is the least computationally expensive, and can still yield good results if the user and source mouths happen to match relatively well, which is not an uncommon occurrence. Such results may be preferable when computational resources are constrained, or the user finds the results to be sufficient. The simple embedding and blending of step (6) comes with more computational costs than the null case, but less computational costs than the GANs based methods of step (7). Step (6) can result

in visual enhancement when compared to the null case, so it may be deemed more appropriate than the null case by the user. The GANs based methods of step (7) are the most computationally intensive. Such methods may be deemed appropriate of the user considers visual realism or greater optionality in the range of outputs to be more relevant than computational efficiency. Notably step (7) may be paired with step (6). Such a pairing may be performed so as to reduce the amount editing to be done by GANs based methods, thus reducing the overall computational burden. This pairing is demarcated by label 3C. GANs which edit the face in its entirety, which here may be employed in step (7), can be used to better blend a mouth into a face. Such GANs may be used to take an unregular image, such as one which has clearly been manipulated, and may be used to make it more photorealistic, and are therefore useful for producing photorealistic mouth deepfakes in the presently considered algorithmic process flow. Finally, step (8) displays and/or exports the result in the desired format. One such format may be live video feed, but a user may have also opted in step (0) to replace a mouth in a single still image rather than a video feed, in which case the resulting image would be displayed. The export of the result would most commonly entail the export of the resulting image and/or video to a standard file format, although this may entail other methods or exporting results. One such alternative export format might entail, but is by no means to be construed to be limited to, the export of an array containing data about

the pixel locations of the points of the mouth segmentation performed in relation to the user's image. Such an export might be done if we anticipated the later use of such data for additional deepfake manipulation of the same image.

[0022] FIG. 4 depicts a simple sketch of the human lips analyzed via facial and/or mouth landmarks. Label 4A denotes the outer lips and label 4B denotes where the upper and lower lips meet. Label 4C denotes one of the landmark points, depicted as a white circle. The landmarks referred to are simply points along the upper and lower lips. The exact location of such landmarks is based on the algorithm which places them. For instance we might choose to design a landmark algorithm which places just 1 landmark point at the estimated center of the mouth, a landmark algorithm which places 2 points – 1 at each corner of the mouth, or even landmark algorithms which place thousands of landmark points. A reduction in computational efficiency is generally accrued as we use more landmarks. These landmarks can be autonomously placed on a user's mouth relatively computationally efficiently via a landmark algorithm .

Label 4D denotes an example error box for the landmark points, such that the landmark points constitute the centroids of their respective error boxes. Label 4E denotes a mouth orientation axis generated from the estimated locations of the

mouth corners landmark points. Such an axis is used to properly orient the deepfaked mouth with respect to the original orientation of the user's mouth. Lines such as the line denoted by label 4E may be generated for both the source mouth and the user mouth, which, when paired with data about which side of the line faces the upper lip and which side faces the lower lip, allows for a relatively efficient and effective alignment of the orientations of both mouths, particularly in instances where both mouths are largely front-facing. Label 4F depicts the background. In the case of landmark segmentation this background becomes the background mask of the mouth. This is similar to the background mask of FIG. 2 denoted by 2B and we shall discuss the case of equivalence below. Landmark algorithms, by autonomously locating the user's mouth, allow us to segment out mouths from images via the generation of segmentation masks in a manner often – yet not always – similar to the segmentation masks depicted in FIG. 2. We may replicate the masks depicted in FIG. 2 through the choice of an appropriate landmark segmentation. In such a replication of FIG. 2, label 4F would denote a background mask equivalent to that denoted by label 2B of FIG. 2. Although landmark segmentation may produce segmentation results identical to those depicted by the polygonal segmentation of FIG. 2, the two segmentation methods are not necessarily equivalent, as FIG. 2 may be obtained via other methods and landmark segmentation may produce different results. For an instance of the latter,

consider the landmark segmentation of 4 points of the mouth, the horizontal corners, and the middle points of the outer rims of the upper and lower lips: such a segmentation would yield a diamond-shaped landmark segmentation. For our purposes it is often helpful to assert this case of equivalence, as we generate landmarks to preserve orientation, so it is more efficient to use those already-calculated points for segmentation.

Mouth landmark identification herein refers to the identification of mouth landmarks as depicted by label 4C of FIG. 4. While a part of landmark segmentation, mouth landmark identification can also provide useful information itself.

Landmark segmentation and mouth landmark identification can be particularly useful in that they allow for mouth segmentation, the high-fidelity estimation of mouth orientation, the high-fidelity estimation of mouth position, and the high-fidelity estimation of mouth size, as well as providing the capacity to determine the location of specific points of the mouth, thereby providing us with information about the structure and/or shape of the mouth in question. Landmark segmentation and mouth landmark identification allow for an efficient deepfake algorithmic process flow by directly providing the information necessary to perform, entirely or in part, steps (1), (2), (4), (5), and (6) of the algorithmic process flow depicted in FIG. 3. Indeed, a relatively high-fidelity and computationally efficient mouth

deepfake can be obtained using these landmark-based methods alone through the selective use of source mouths. For instance, we may choose a source mouth which matches the gender and skin tone of the user and simply use data provided by the landmark algorithm to segment the source mouth, resize it, orient it according to the orientation of the user's mouth (the discussion concerning 4E and the utilization of the axis connecting the corners of the mouth is one such method for the alignment of the orientation of the two mouths). We will term mouth deepfakes which rely mainly upon landmark-based processes, such as in this instance, "landmark-centric deepfakes". Landmark-centric deepfakes may be altered and/or enhanced via other methods such as, but not limited to, the GANs-based embedding of step (7) of FIG. 3 for visual enhancement, but most of their structural alterations relating to any type of pixel positional data (such as we use when determining size, shape, and orientation) is to be understood to be retrieved from landmark-based algorithms. The case of deepfake construction from solely landmark-based methods will be termed "landmark-only deepfakes". The two instances are largely similar, so we will frame the present discussion in terms of the more robust landmark-centric deepfakes. As landmark-only deepfakes are simply the boundary case of landmark-centric deepfakes without alteration and/or enhancement via other methods (therefore landmark-centric deepfakes may become landmark-only deepfakes while the reverse is not true), they definitionally

can only perform as well as landmark-centric deepfakes. Additionally, the more robust landmark-centric deepfakes are relevant to real world use cases, where it is undesirable to restrict ourselves to the boundary case unless absolutely necessary. Therefore, the discussion of the efficacy of landmark-based deepfakes will be framed in terms of landmark-centric deepfakes.

Landmark-centric deepfakes can yield relatively high-fidelity results if features such as lip and/or mouth structure and/or color and the skin tones match in a manner which suits the intended application. What the user finds suitable is highly subjective and reliant largely upon the opinions of the user in question. We have found that the use of slightly lighter skin tone source mouths can produce a result which is relatively aesthetically appealing to some users, with the lighter skin tone of the source face producing a spotlight-like effect, brightening, and highlighting the area around the users' mouth. In such instances the landmark-centric deepfake mouth may even appear to be that of the users with a simple brightness effect applied around the mouth. To the contrary, changing the skin tone too much or to a darker shade can be perceived negatively. In the case of lip alteration, in which a user may wish to explore the effects of lip-focused cosmetic procedures, landmark-centric deepfakes produce largely sufficient results. Landmark-centric deepfakes also allow for some things not easily replicated in other algorithms. They may for instance, be directed to change their basic structure or size in accordance with the

user's mouth movement. Since both the user mouth and source mouth may be analogously landmarked, we may create intuitive interfaces for the user to be able to alter the landmark-centric deepfake. For instance, we have produced landmark-centric deepfakes which alter in size based on how large or small the user makes their own mouth (larger or smaller in the image taken of them). Thus, the user may simply widen their mouth through actions such as smiling to make the deepfake mouth larger or make their mouth smaller via actions such as puckering to reduce the size of the mouth. To the contrary, alternative methods of producing deepfake mouths, such as GANs, may select the size of the deepfake autonomously and not allow the user much control. While GANs may technically allow for the control of mouth size via the manipulation of vectors in their representation spaces; this can entail intervention by an experienced operator of the algorithm or require additional steps to allow the user to alter the deepfake in as intuitive a manner as can be done with landmark-centric deepfakes, which can entail a dramatic decrease in computational efficiency. Landmark-centric deepfakes may also be used in anomalous cases. Some cosmetic procedures may produce unnatural or otherwise uncommon results, with some procedures performed specifically to look unnatural or uncommon. By either autonomously or manually labelling landmark positions along some anomalous source lips, we can use those positions to place the anomalous source lips onto the user's mouth. To the contrary, GANs are generally

reliant upon the representation spaces produced by some given previously constructed dataset. This means they encounter difficulties in anomalous cases. So, if some new style of lip alteration suddenly came into fashion, we would first need to gather hundreds, thousands, or even tens of thousands of examples of such lips and then retrain the GANs or train a new GANs with that new data included. With landmark-centric deepfakes, we would only need to label the points where we should expect the landmarks to appear, and this only needs to be done on a single example to make landmark-centric deepfakes operational. For instance, consider the case of a 4-landmark model: 2 on corners of the mouth, 1 on the middle of the upper lip, 1 on the middle of the lower lip. Many types of images – not just anomalous lips – could be overlaid on the user's mouth by simply labelling 4 points of the images as landmarks. And we only need 1 example of some set of anomalous lips to produce landmark-centric deepfakes of that set of lips, a clear advantage over GANs-based methods. Landmark-centric deepfakes are also relatively computationally efficient compared to more advanced GANs-based deepfakes.

The use of landmark-centric deepfakes has limitations.

Landmark-based algorithms are also more effective in detecting more obvious boundary regions, and may encounter difficulty in determining the location of less obvious boundaries. This means that they may detect the outer edges of the lips

relatively efficiently, as lips strongly contrast with surrounding skin . But if we wished to determine the boundary between the two lips then the employed landmark-based algorithms may lose efficacy to the degree where other algorithms become favorable. We have, for instance, generally observed selective region parsing algorithms, such as those we will discuss in the context of FIG. 6, to be more precise and more efficient at determining such boundaries (although such results are highly dependent upon the design of the respective algorithms). An important limitation is the lack of ability to move the mouth's individual parts in a photorealistic manner. While the mouth as a whole can be moved across an image, video and/or a face therein with relatively high fidelity, actions such as moving the lips apart can produce imprecise results. For such alterations we resort to more complex GANs-based mouth deepfakes.

The various segmentation models discussed in the present work can be similarly used to construct what we will term to be "self-centric deepfakes": deepfakes where all pixel positional information is determined solely from the model in question yet may be enhanced or altered via other methodologies. Landmark-centric deepfakes are the class of self-centric deepfakes constructed from landmark-based models. We will limit our discussion of self-centric deepfakes to the landmark model case, as landmark-centric deepfakes are the class of self-

centric deepfakes most applicable to the real-world applications of the considered mouth deepfake algorithm.

[0023] FIG. 5 depicts another possible methodology for human face segmentation, with multiple major elements of the human face isolated, identified, and assigned a label. This contrasts FIG. 2, in which only the mouth is segmented out. The mouth-only segmentation of FIG. 2 is adequate for many scenarios and is more computationally efficient. However, the full-face segmentation of FIG. 5 may be more appropriate in certain scenarios. Such scenarios may include, but are not limited to, when: we need to do a GANs based embedding and blending of the mouth into the user's face as depicted in step (7) of FIG. 3 and discussed in [0020]; we do a GANs based adjustment of the source mouth as depicted in step (3) of FIG. 3 and discussed in [0020]; we want to do a simple embedding and blending into the user's face as depicted in step (6) of FIG. 3 and discussed in [0020]; or we want to apply a corrective morphing as depicted in step (4) of FIG. 3 and discussed in [0020]. In such scenarios we might apply data from the full-face segmentation of FIG. 5 in order to construct a more ideal output, per some provided criteria. For instance, if we are attempting to make a visually appealing output such that the

user perceives themselves to look more visually appealing with the new mouth and teeth, then we might rely on the assumption of equating human visual appeal and face symmetry and proportionality, and therefore we would use the full-face segmentation data to construct and place the deepfaked mouth in such a way as to maximize symmetry and proportionality in regard to that user's nose, jawline, or other elements of their face. Label 5A points to the numerically labelled and identically constructed x-axis and y-axis of the upper right full-face segmentation. The sample of FIG. 5 portrays segmentation images of pixel size 512x512 or 262,144 pixels in total. Importantly 512 is equivalent to 2 to the 9^{th} power, thus making 262,144 equivalent to 2 to the 18^{th} power. Hence the axes denoted by 5A, as well as all the other axes, denote pixel values 1 to 512, with pixel x=1, y=1 being the lower left pixel of each segmentation and pixel x=512, y=512 being the upper right of each segmentation. Images which are multiples of some power of 2 are more easily and more ably injected into a neural network based process flow. Within the context of the neural network we can then treat the images as a Dx(2^n)x(2^n) mathematical tensor, where (2^n) is 2 to the n-th power (n=9 for the example of FIG. 5), and where D is some tensor which may relate to the structure of our input image or the structure of our neural network; in the simplest case D can reduce to a 0-Tensor (a scalar), such as the integers 1 or 3; and in some more complex cases we may use a 2-Tensor such as a 512x512 matrix which may

represent a image the neural network uses to manipulated our processed image more capably. These numerically specific examples are only to further elucidate the underlying algorithmic mechanisms and by no means cover the full scope of possible scenarios. So, while our algorithm is not restricted to images of this size, it is often practical to resize the images to such power of 2 scales at one or more points in our algorithmic process flow.

Label 5B denotes the segmented human mouth. In this particular depiction the upper lip, lower lip, and inner mouth (such as teeth in the case of a smile) are each isolated as individual regions. Such separation of the upper lip, lower lip, and inner mouth, can help us construct a more representative deepfake of the user's mouth. A more detailed instance of such tripartite mouth segmentation and its relevance to the present invention will be discussed in the context of FIG. 6.

Label 5C denotes a slanted dashed line through one of the full-face segmentations. This slant is to denote and highlight the fact that faces will not always be straightly aligned, it may be the case that we need to correct such misalignment. Our algorithm can generally account for such misalignments without additional correction during standard use. But in the instance of extreme misalignment, such as in the instance of a child sporadically moving their head about, it may be necessary to devote additional computational resources to correct the misalignment issue. In such an instance we can use other segmented parts of the face to help

increase the probability of proper alignment. From the face on which the dashed line denoted by 5C is placed, we may, for instance, use an approximately bisecting line of the width of the nose, a line bisecting the width of the nose, the approximate midpoint between the inner edges of the eyes, the approximate midpoint between the inner edges of the eyebrows, the approximate midpoint between the ears, and the approximate midpoint of the chin to attempt to estimate the horizontal location of the mouth and we may use the position of the jawline and the position of the lower edge of the nose to attempt to estimate the approximate vertical location and size of the mouth. Given the relative similarity of human facial proportions, such estimations can be quite accurate, even in the absence of any information about the mouth itself.

We will now discuss other relevant instances where full-face segmentation can aid in proper alignment in the context of labels 5D and 5E, with a particular focus on injuries such as those which may be suffered by burn victims which may destroy the lips and deformities of the mouth such as cleft palate. In such instances the mouth segmentation may fail in the full-face segmentation algorithm, but the mouth-only segmentation methods of FIG. 2 and FIG. 4 might fail more completely, necessitating the processes detailed below.

Label 5D denotes what might be seen (and is in fact seen) in the instance of a deep learning based full-face segmentation on a face which is not showing any lips at

all, with the label denoting a depiction of the segmented face and a close-up of the mouth region. In such a no-lip case the mouth region can largely be discerned as a slit in the face or otherwise. Such a no-lip appearance is achievable by many humans simply by pulling the lips into the mouth. Such a no-lip case is particularly pertinent in regard to injuries which may result in the destruction of the lips, such as injuries which may be suffered by severe burn victims. The close-up exemplified by 5D relates that a deep learning full-face segmentation algorithm may be able to estimate where the lips should be – and indeed this occurs in real segmentations of this type – but this leaves us relatively unable to estimate the real size and full position of the mouth. Deep learning, machine learning, and artificial intelligence algorithms are largely statistical algorithms, so the anomalous nature of the no-lips case may result in unexpected and unpredictable distortions in the applicability of the algorithms themselves and may render them particularly ineffective in precision estimations in anomalous cases. We can, however, exploit the relative similarity of human face proportions in order to estimate the proportions and location of the mouth in question. Using other segmented parts of the face, we may use the various other proportions and locations of facial structures (such as the examples provided earlier in this section in the context of label 5C) to estimate the proportions and location of the mouth and lips with a great deal of accuracy.

Label 5E denotes what might be seen (and is in fact seen) in the instance of a deep learning based full-face segmentation on a face which suffers from a cleft palate and a close-up of the mouth and lips of that face. Since deep learning, machine learning, and artificial intelligence algorithms are statistical they can yield poor estimates when dealing with anomalous cases. The representation space of a deep learning algorithm which might produce the image labelled by 5E is unlikely to contain much information about cleft palates, so the algorithm attempts to fit the affected lip into the context of the standard human lip that it is familiar with. Such a process can result in the distortions such as those exemplified in FIG. 5 label 5E: a deep learning full-face segmentation algorithm might remove the cleft connecting the mouth to the nose but may leave distortions such as the wave-like distortion denoted by 5E. In such an instance we may either remove the mouth and lip segmentation regions from the image entirely and reassess the location and proportions of a reconstructed mouth based on other details of the face, as done previously. Alternatively, we may use the location and size approximated by the algorithm but alter the structure to that of a typical mouth in the deepfake output. Some combination of these may also be applied. The GANs based methods of step (7) of FIG. 3 may then be applied to edit the distorted skin to replace the cleft between the mouth and the nose with significantly more normal-appearing skin.

The methodologies discussed in the context of FIG. 5 may be regarded as taking place in a number of steps of the algorithmic process flow depicted in FIG. 3. In particular, the process of determining whether the algorithm is dealing with the no-lips or cleft palate case may be done in step (1) of FIG. 3 or may be manually input by the user as per step (0) of FIG. 3 or some combination thereof.

[0024] FIG. 6 depicts the process of the selectively editing teeth via tripartite mouth segmentation. FIG. 6 will show more clearly how we may further segment the mouth itself into its constituent components and utilize this for a practical and selective editing of the constituent parts of the human mouth, not just replace the entire mouth. Label 6A denotes a sketch of a hypothetical user's mouth after extraction from the user's face, so the area outside the lips can be considered to be the background mask. The mouth of 6A is missing 2 top teeth. Label 6B denotes a potential deep learning based segmentation of this mouth. Label 6C denotes the segmentation region of the upper lip, shown in gray in FIG. 6. Label 6D denotes the segmentation region of the inner mouth, which in this case is specifically defined as the sum of the mouth between the upper and lower lip. So the region denoted by 6D would include the teeth, gums, and other parts of the inner mouth as seen in the sketch denoted by 6A. Label 6E denotes the segmentation region of the lower lip, shown as the black region of the lips. The

region outside the lips simply remains the background mask. Using this 3 region segmentation we can now replace the previous teeth-deficient inner mouth denoted by 6A with an inner mouth which contains a full set of teeth. Such a full set of teeth can be chosen from a source such as a stock image model, or the user may opt to edit their own mouth, treating it as the source mouth as described in [0021] in the context of FIG. 3. Given the ideal case of near-perfect symmetry present in this particular sketch example, the desired editing may be achieved via the simple editing of step (4) of FIG. 3 by simply taking the mirror of the full teeth side and superimposing it over the missing teeth side. It is to be noted, however, that real mouths are rarely perfectly symmetrical in such a manner, so in the real mouth scenario we may see dramatic improvements via the use of the mouth-focused GANs of step (3) of FIG. 3. Perhaps steps (3) and (4) of FIG. 3 in conjunction may exploit the approximate symmetry of a real mouth, that is a rough mirrored image could be first be produced vi step (4) which can then be run through step (3). Such a conjunction would have the advantage of saving the algorithm from having to construct replacement teeth using GANs – rather the GANs would only need to adjust the mirrored image – thus allowing us to increase the computational efficiency of the overall algorithm.

[0025] Generative Adversarial Networks (GANs) are complex deep neural network models which may be used to autonomously generate images from mathematical objects known as latent vectors or from source images, or via a combination of such factors. Such GANs allow for the construction of complex and separable high-dimensional latent vector spaces consisting of basis vectors which may be associated with some real emergent property. Of particular relevance is the ability of GANs to effectively construct and manipulate highly realistic images of the human face, the constituent components of the human face, amongst other objects. In recent years such models have been able to reconstruct the human face to the point of photorealism. A layperson unaware of the characteristic distortions associated with a particular GAN model may be entirely unable to distinguish such images from real photos of human faces. GANs may be applied in 3 major ways in the present system: generation, embedding, latent vector space traversal. Generation allows the for the creation of new mouths for the user to try on. Embedding allows for the enhanced blending of an extant mouth into the user's face. Latent vector space traversal allows us to edit mouths, either the user's original mouth, which we can then proceed to deepfake back onto the user or edit generated mouths.

[0026] The segmentation models presented herein are further utilized in a novel fashion for the creation of novel datasets which can be utilized by GANs for the creation of mouth-focused GANs discussed in step (2) of the algorithmic process flow of FIG. 3. We utilize mouth segmentation models in order to isolated the human mouth in human face imagery for the creation of large mouth-only datasets upon which GANs may be trained. For instance, we may by the presented method use the same datasets used to train full-face GANs to train mouth-only GANs without the need to manually isolate the human mouth in each image. GANs trained on the human mouth alone are dramatically more capable of editing the human mouth than GANs trained on the full face. Suppose, for instance, that a 512-dimensional full-face GANs latent space contained only 16 basis vectors which manipulated properties of the human mouth (a not-atypical situation). Then, by constructing a 512-dimensional mouth-only GANs, nearly all 512 basis vectors can be expected to alter the mouth in some manner (barring those which alter properties such as brightness, light direction, etc. or pick up other image artifacts). Or we may alternatively utilize a 32-dimensional mouth-only GANs to achieve superior mouth-alteration (a 16-dimensional GANs might contain basis vectors related to properties such as brightness, light direction, etc., hence the use of a 32-dimensional GANs) as compared to the full-face GANs. GANs are exceedingly

computationally expensive models so such a dimensionality reduction results in a dramatic improvement in computational efficiency.

Additionally, by using the same segmentation model that we use to create mouth datasets in order to run the algorithm we can ensure that the mouth-only GANs is well-equipped to process that particular type of segmentation model.

[0027] Through our system, GANs may be applied selectively to the human mouth. Through human mouth segmentation similar to that discussed in the prior explanation of FIG. 2, we may isolate the mouth and build a generative adversarial network upon only the mouth. Such novel isolation has the advantage of applying a potentially large latent space to the mouth alone. According to the current standard of untargeted GANs face manipulation, the latent vectors of the GAN must describe the entirety of the human face or the human form. An identical GAN applied to the entirety of the human face or the entirety of the human form would not be capable of capturing and representing as many features of the human mouth as it would when applied to the mouth alone. Additionally, we are able to gain a much greater representation of the human mouth with a much smaller GAN model when the mouth is isolated. Such reductions may result in up to and exponential improvement in efficiency in our use of GANs for human mouth

manipulation, yielding large computational improvements. Through precision targeting of the mouth, we enable the latent space of the GAN to more accurately and completely describe the human mouth far more efficiently.

GANs which edit the face in its entirety have been explored in the field of artificial intelligence. The present invention also addresses a number of major problems with – and associated with – such full-face GANs: the lack of control of mouth editing; the high prevalence of mouth distortions in even the most advanced contemporary models; the over-interpolation of mouth-controlling basis vectors with basis vectors unrelated to the mouth itself; and the inability of such models to successfully debias mouth editing. For instance, the mouth-controlling basis vectors of even the contemporary world's most advanced GANs have been found to contain undue racial, agist, sexist, and ethnic biases – biases which are presently contrary to various laws of the United States and its many jurisdictions – in regard to the structure of the mouth. For instance, many full-face GANs inappropriately and unduly correlate thicker lips with non-white individuals, frowns with older individuals and males, timid expressions with females, and yellow teeth with particular ethnic minorities. This means that users seeking to edit their mouths and teeth with such GANs are also quite likely to see changes in skin color, ethnic group, facial bone structure, or even gender, a phenomenon which clearly renders such models inoperable when it comes to the face-preserving GANs based mouth

associated with the present invention, which, by contrast, can edit a user's mouth on their own otherwise-unaltered face.

It is furthermore mathematically impossible to achieve the same degree of mouth editing options using an full-face GANs as it is using the mouth-focused GANs of the present invention. This allows the present invention to edit user mouths far more computationally efficiently, and therefore more seamlessly and in a more cost-effective manner.

[0028] Some distinguishing factors of the present invention will now be mentioned.

Deepfaking of the human mouth, and—quite essentially—only the human mouth. This focus on the mouth and only the mouth allows for the significantly more efficient use of computational resources. Crucially, full face deepfakes are too computationally burdensome for many modern devices, and, on devices actually capable of running such models, are exceptionally slow to run. This computational inefficiency also restricts the use of real time full face deepfakes on all but the most powerful of modern computational devices. Our focus on the mouth allows for the use of models 100 to 1000 times more efficient than many full face deepfake models, while retaining a similar photorealistic result. This resulting

enhanced efficiency also allows us to create such deepfakes in real time without the use of sophisticated computational hardware, which allows this technology to used more broadly and more easily. Restricting our focus to the mouth also prevents the undue correlation or interpolation of basis vectors in applied deep neural networks. A deep neural network which models the deepfaking of the entire human face may, for instance, erroneously correlate the opening of the eyes with the closing of the mouth. Such unintended basis vectors correlations are common in the latent vector spaces of deep neural networks. Our models are less likely to produce such undue correlations.

Another key distinguishing feature is the use of autonomous systems for such photo editing tasks. The human mouth has been edited in various forms of imagery for hundreds, if not thousands, of years. In recent decades photo editing software has been applied to the task. But this digital editing still requires manual input and/or adjustment by a human. The present invention fully automates the task. Such automation also allows for the creation of real time video feeds with created from such altered images, a result which is not similarly achievable with human edited photographs. Current experiments suggest that – even in the case of the alteration of mouth in a single still image – the algorithms presented herein, when optimized, may be run for as low as one one-thousandth to one one-hundredth or 0.1% to 1% of the cost of comparable human labor on a per-hour basis.

Another key distinguishing feature of the present invention is the application of artificial intelligence, machine learning, and deep learning algorithms for such a task. The utilization of Generative Adversarial Networks (GANs), a class of artificial intelligence algorithms, for the outlined tasks, is similarly novel.

The combining of these distinguishing features for the task of human mouth alteration is also in itself unique.

CLAIMS

What is claimed is:

1) An artificial intelligence utilizing algorithm for editing and replacement of the human mouth comprising:

1a) the real-time editing and replacement of the human mouth.

1b) enhanced mouth deepfake efficiency by focused editing of only the human mouth.

1c) identity-preserving mouth deepfake creation which only targets a user's mouth, with the capability of leaving the rest of the user's face otherwise unaltered.

2) The utilization of the herein defined self-centric deepfakes, with a particular focus on landmark-centric deepfakes, for the efficient deepfaking of the human mouth comprising:

2a) novel methodologies for creating deepfakes of the human mouth from the pixel positional data extracted through segmentation models.

2b) novel methodologies for creating deepfakes from the pixel positional data extracted through landmark models in conjunction with landmark segmentation.

3) The novel use of segmentation algorithms for the enhancement of the process of creating deepfakes of the human mouth and the enhancement of associated processes comprising:

3a) novel methodologies for the use of mouth segmentation for the construction of datasets on which to train advanced generative adversarial networks on the isolated human mouth.

3b) novel methodologies for the integration of human mouth segmentation into the algorithmic process flow for the creation of mouth deepfakes.

4) The novel use of generative adversarial networks for the targeted generation, editing, manipulation, and/or embedding of the targeted and/or isolated human mouth comprising:

4a) the novel use of generative adversarial networks trained only on the human mouth.

4b) the conjunction of the novel methodology of claim 4a with the embedding of mouths generated by mouth-only generative adversarial networks into the human face through the utilization of full-face generative adversarial networks.

4c) the novel use of generative adversarial network associated latent vector space traversal for the: editing of, manipulation of, avoidance of undesired interpolation of mouth-related basis vectors with basis vectors unrelated to the manipulation of, and/or the debiasing of deepfakes of, the targeted and/or isolated human mouth.

FIG. 1

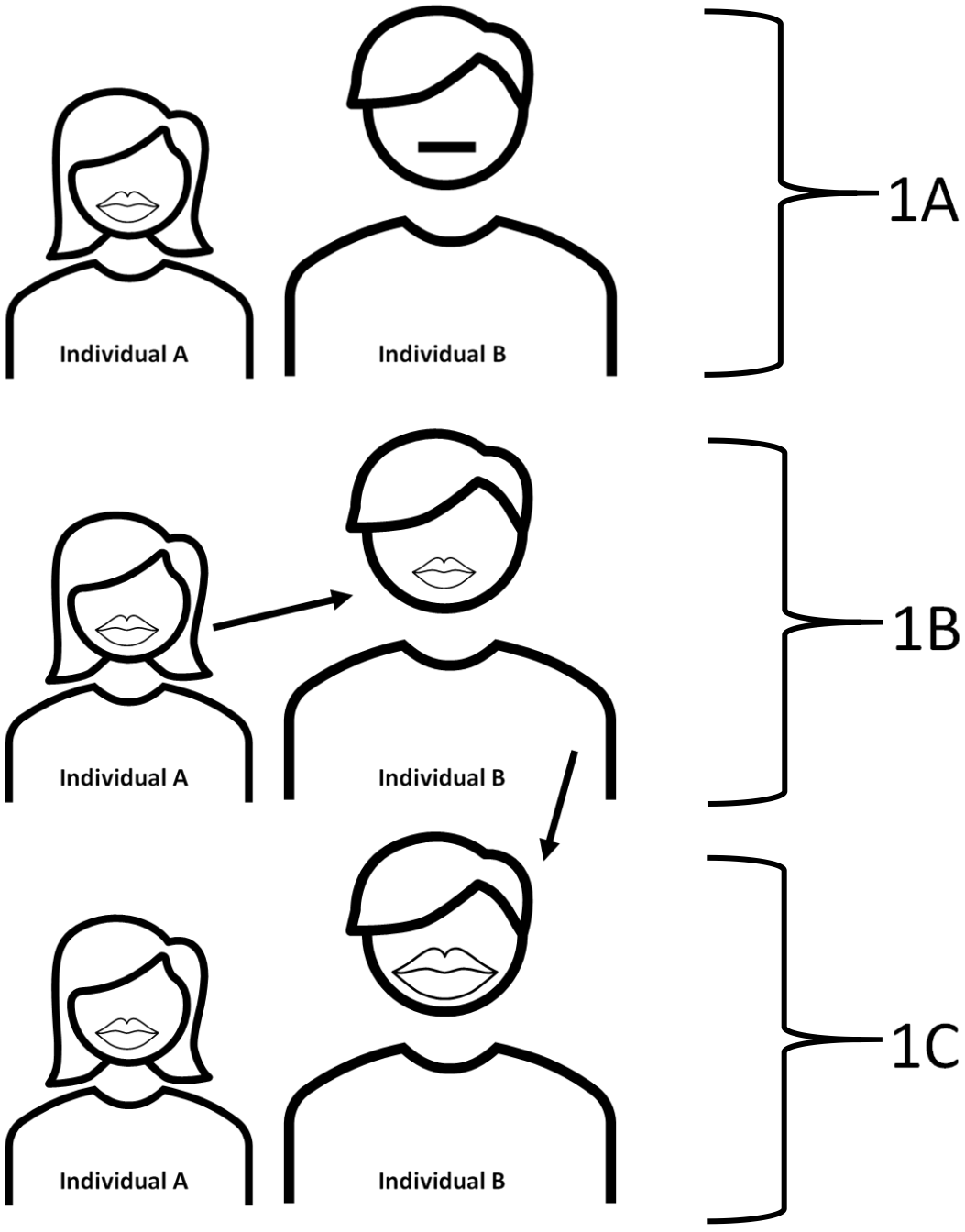
Individual A
Individual B
1A
Individual A
Individual B
1B
Individual A
Individual B
1C

FIG. 2

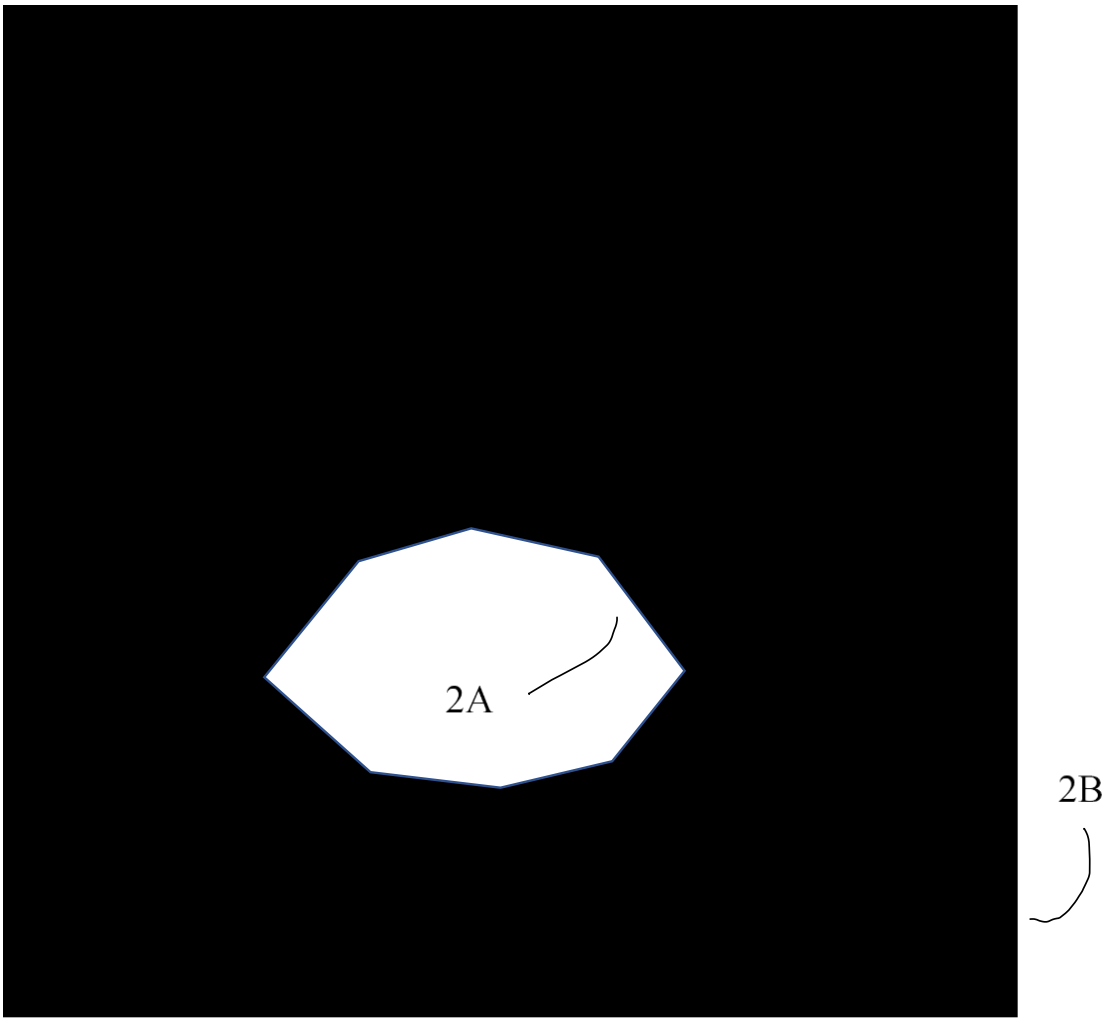

FIG. 3

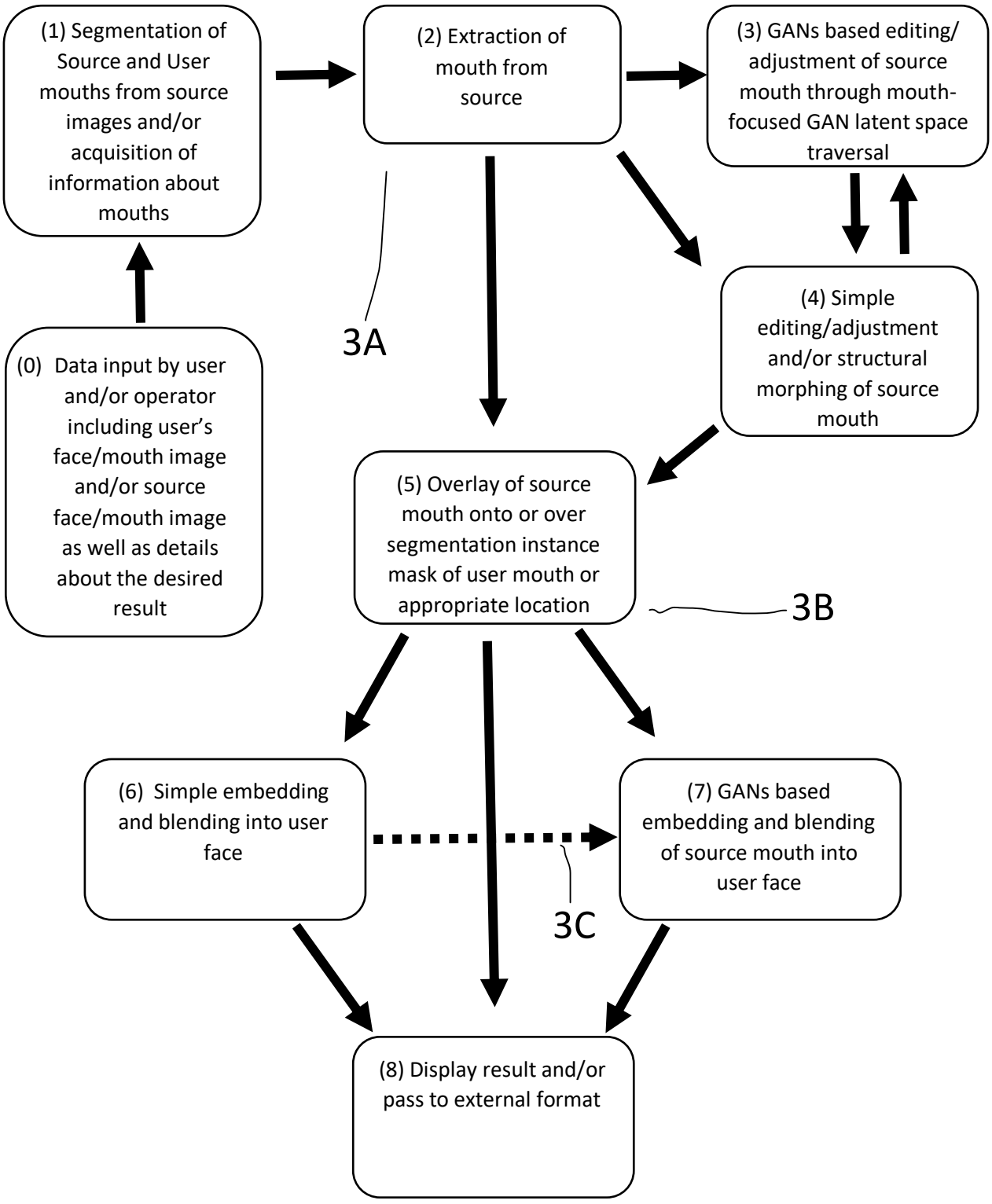
(1) Segmentation of Source and User mouths from source images and/or acquisition of information about mouths
(2) Extraction of mouth from source
(3) GANs based editing/ adjustment of source mouth through mouth-focused GAN latent space traversal
(4) Simple editing/adjustment and/or structural morphing of source mouth
(0) Data input by user and/or operator including user's face/mouth image and/or source face/mouth image as well as details about the desired result
3A
(5) Overlay of source mouth onto or over segmentation instance mask of user mouth or appropriate location
3B
(6) Simple embedding and blending into user face
(7) GANs based embedding and blending of source mouth into user face
3C
(8) Display result and/or pass to external format

FIG. 4

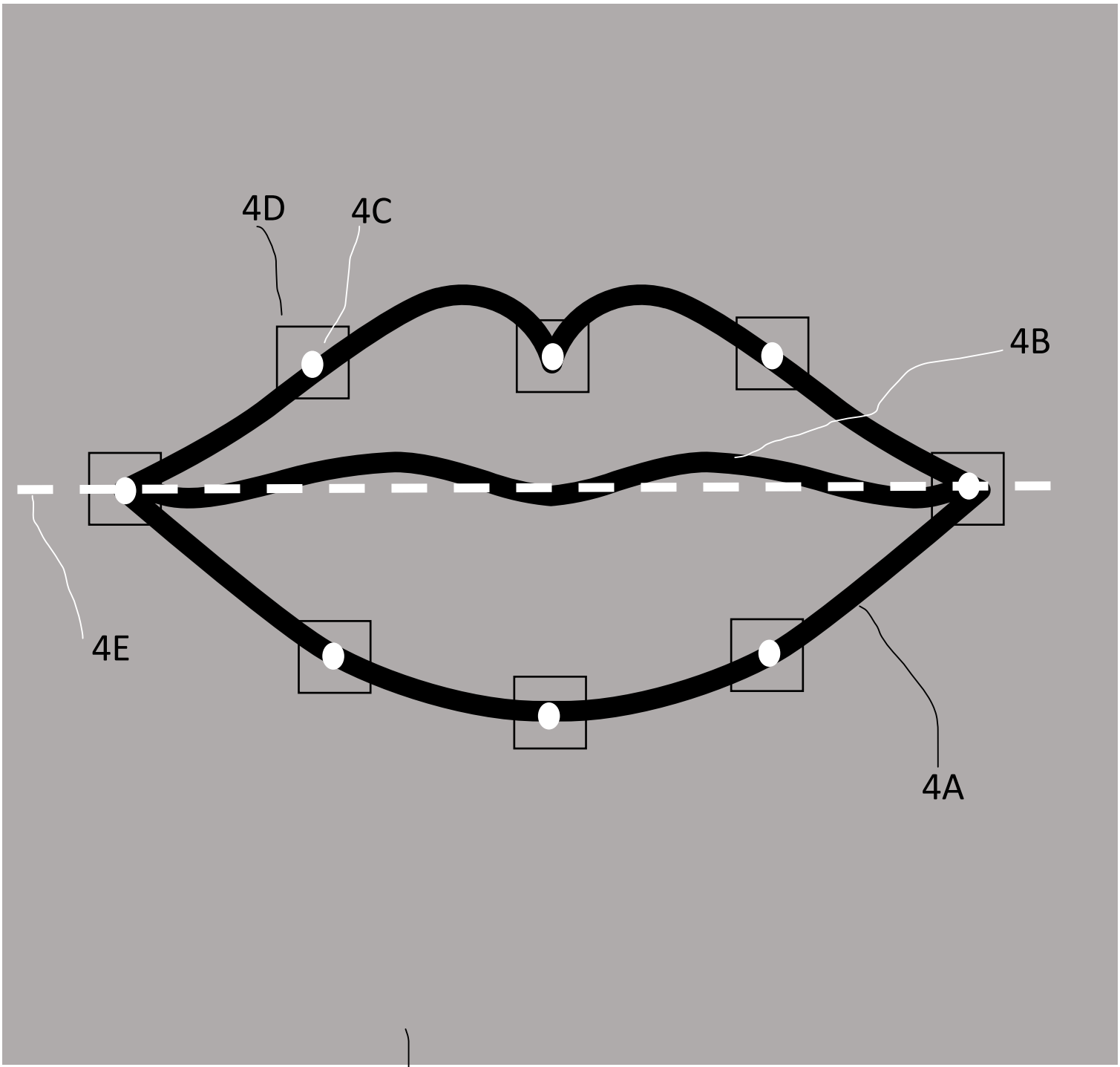

FIG. 5

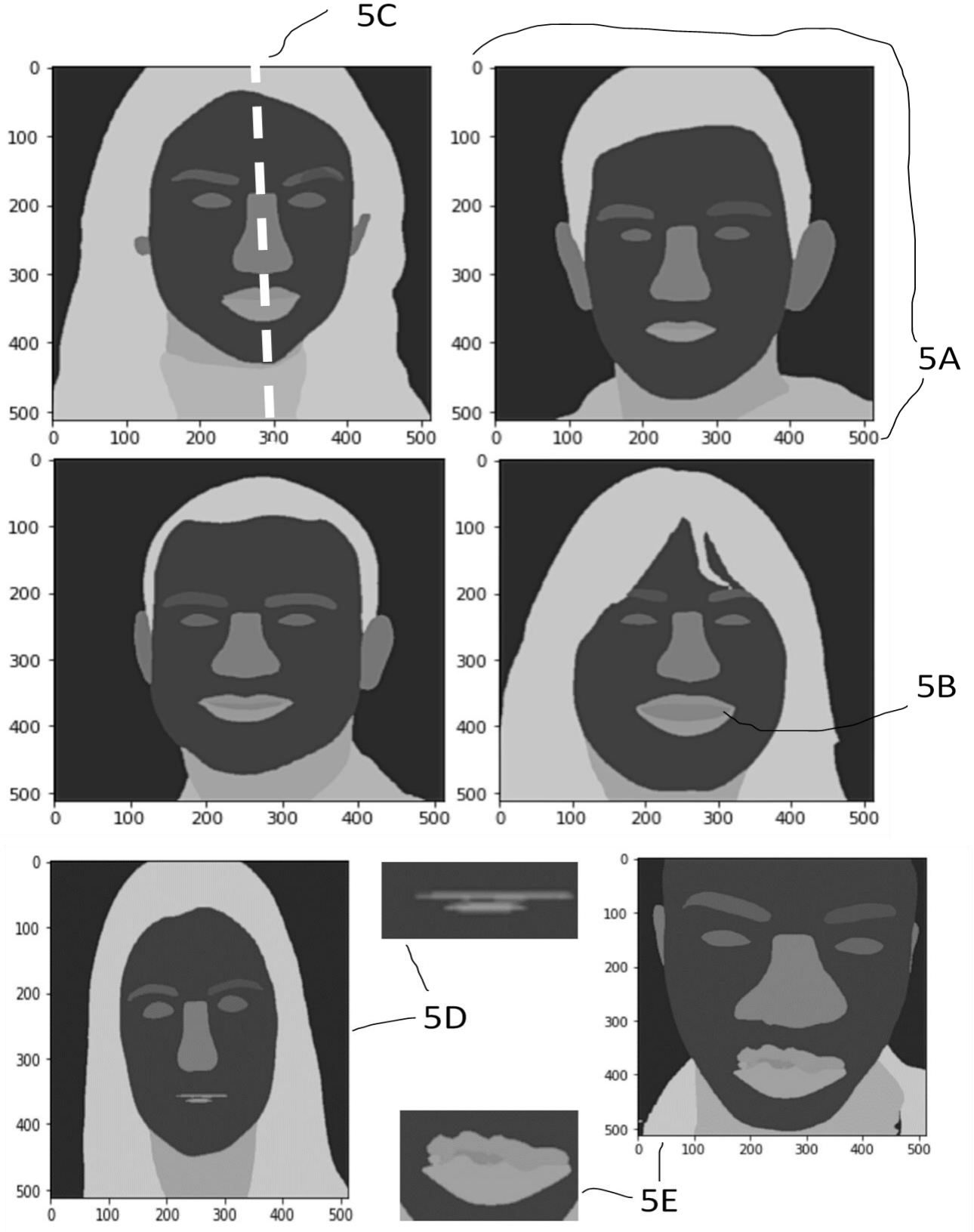

FIG. 6